

Riemannian stochastic quasi-Newton algorithm with variance reduction and its convergence analysis

Hiroiyuki Kasai*

Hiroiyuki Sato[†]

Bamdev Mishra[‡]

April 13, 2017

Abstract

Stochastic gradient algorithms have recently gained significant attention for minimizing the average of a large, but finite number of loss functions. The present paper proposes a Riemannian stochastic quasi-Newton algorithm with variance reduction (R-SQN-VR). We tackle with the key challenges of averaging, adding, and subtracting multiple gradients by exploiting notions of retraction and vector transport. We present a global convergence analysis and a local convergence rate analysis of R-SQN-VR under some natural assumptions. The proposed algorithm is applied to the Karcher mean computation on the symmetric positive-definite manifold and low-rank matrix completion on the Grassmann manifold. Exhaustive experiments reveal the superior performances of the proposed algorithm in comparison with the Riemannian stochastic gradient descent and the Riemannian stochastic variance reduction algorithms.

1 Introduction

Let $f : \mathcal{M} \rightarrow \mathbb{R}$ be a smooth real-valued function on a Riemannian manifold \mathcal{M} . The problem under consideration in the present paper is the minimization of the *expected risk* of f for a given model variable $w \in \mathcal{M}$ taken with respect to the distribution of z , i.e., $\min_{w \in \mathcal{M}} f(w)$, where $f(w) = \mathbb{E}_z[f(w; z)] = \int f(w; z) dP(z)$ and z is a random seed representing a single sample or set of samples. When given a set of realizations $\{z_{[n]}\}_{n=1}^N$ of z , we define the loss incurred by the parameter vector w with respect to the n -th sample as $f_n(w) := f(w; z_{[n]})$, and then the *empirical risk* is defined as the average of the sample losses:

$$\min_{w \in \mathcal{M}} \left\{ f(w) := \frac{1}{N} \sum_{n=1}^N f_n(w) \right\}, \quad (1)$$

where N is the total number of the elements. This problem has many applications that include, to name a few, principal component analysis (PCA) and the subspace tracking problem [1] on the Grassmann manifold, which is the set of r -dimensional linear subspaces in \mathbb{R}^d . The low-rank matrix/tensor completion problem is a promising example of the manifold

*Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan (kasai@is.uec.ac.jp).

[†]Department of Information and Computer Technology, Tokyo University of Science, Tokyo, Japan (hsato@rs.tus.ac.jp).

[‡]Core Machine Learning Team, Amazon.com, Bangalore, India (bamdevm@amazon.com.)

of fixed-rank matrices/tensors [2, 3]. The linear regression problem is also defined on the manifold of the fixed-rank matrices [4].

Riemannian gradient descent requires the *Riemannian full gradient* estimation, i.e., $\text{grad}f(w) = \sum_{n=1}^N \text{grad}f_n(w)$, for every iteration, where $\text{grad}f_n(w)$ is the *Riemannian stochastic gradient* of $f_n(w)$ on the Riemannian manifold \mathcal{M} for n -th sample. This estimation is computationally heavy when N is extremely large. A popular alternative is *Riemannian stochastic gradient descent* (R-SGD) that extends *stochastic gradient descent* (SGD) in the Euclidean space [5]. Because this uses only one $\text{grad}f_n(w)$, the complexity per iteration is independent of N . However, similarly to SGD [6], R-SGD requires *decaying step-size* sequences which start with a large step-size and decrease with iterations to guarantee its convergence, but this causes a slow convergence.

Variance reduction (VR) methods have been proposed recently to accelerate the convergence of SGD in the Euclidean space [7–11]. One distinguished feature of them is to calculate a full gradient estimation periodically, and to re-use it to reduce the variance of noisy stochastic gradient. However, because all previously described algorithms are *first-order* algorithms, their convergence speed can be slow because of their poor curvature approximations in *ill-conditioned* problems. One promising approach is *second-order* algorithms such as stochastic *quasi-Newton* (QN) methods using Hessian evaluations [12–15]. They achieve faster convergence by exploiting curvature information of the objective function f . Furthermore, addressing these two acceleration techniques, [16] and [17] propose a hybrid algorithm of the stochastic QN method accompanied with the VR method.

Examining the Riemannian manifolds again, many challenges on the QN method have been addressed in deterministic settings [18–20]. The VR method in the Euclidean space has also been extended to Riemannian manifolds, so-called R-SVRG [21, 22]. Nevertheless, the second-order stochastic algorithm with the VR method has not been explored thoroughly for the problem (1). To this end, we propose a Riemannian stochastic QN method based on L-BFGS and the VR method. Our contributions are three-fold;

- We propose a novel (and to the best of our knowledge, the first) Riemannian limited-memory QN algorithm with a VR method.
- Our convergence analysis separately deals with a global convergence and a local rate of convergence analysis. This analysis is typical for batch algorithms on Riemannian manifolds [23]. Our assumptions for the latter are imposed only in a local neighborhood around a minimum, which are milder and natural. Consequently, our analysis is expected to be applicable to wider varieties of manifolds.
- We present the convergence analysis of the algorithm with *retraction* and *vector transport* operations instead of the more restrictive *exponential mapping* and *parallel transport* operations. This makes the algorithm appealing for practical problems.

The specific features of the algorithms are two-fold;

- We update the curvature pair of the QN method every outer loop by exploiting full gradient estimations in the VR method, and thereby capture more precise and stabler curvature information without relying on unrealizable stochastic gradient estimations. This avoids additional sweeping of samples required in the Euclidean stochastic QN [15] and additional gradient estimations required in the Euclidean online BFGS (oBFGS) [12, 13, 24].

- Compared with a simple Riemannian extension of the QN method, a noteworthy advantage of its combination with the VR method is that, as revealed below, frequent transportations of curvature information between different tangent spaces, which are inextricable in such a simple Riemannian extension, can be reduced drastically by the VR algorithm. This is the special benefit of the Riemannian hybrid algorithm, which does not exist in the Euclidean case [16, 17]. More specifically, the update of curvature information and the calculation of *second-order modified Riemannian stochastic gradient* are performed uniformly on the tangent space of the outer loop.

The paper is organized as follows. Section 2 presents details of our proposed R-SQN-VR. Section 3 presents two theorems of the convergence analysis. In Section 4, numerical comparisons are explained with R-SGD and R-SVRG on two problems, with results suggesting the superior performances of R-SQN-VR. The proposed R-SQN-VR is implemented in the Matlab toolbox Manopt [25]. A brief explanation of optimization on manifold, the concrete proofs of theorems, and additional experiments are provided as supplementary material.

2 Riemannian stochastic quasi-Newton algorithm with variance reduction (R-SQN-VR)

We assume that the manifold \mathcal{M} is endowed with a Riemannian metric structure, i.e., a smooth inner product $\langle \cdot, \cdot \rangle_w$ of tangent vectors is associated with the tangent space $T_w\mathcal{M}$ for all $w \in \mathcal{M}$ [23]. The *norm* $\|\cdot\|_w$ of a tangent vector is the norm associated with the Riemannian metric. The metric structure allows a systematic framework for doing optimization over manifolds. Conceptually, the constrained optimization problem (1) is translated into an *unconstrained* optimization problem over \mathcal{M} . Consequently, notions such as the Riemannian gradient (first-order derivatives of an objective function), tangent space, and moving along a search direction have well-known expressions for a number of manifolds.

2.1 R-SGD and R-SVRG

R-SGD: Given a starting point $w_0 \in \mathcal{M}$, the R-SGD algorithm produces a sequence $(w_t)_{t \geq 0}$ in \mathcal{M} that converges to a first-order critical point of (1). Specifically, the R-SGD algorithm updates w as $w_{t+1} = R_{w_t}(-\alpha_t \text{grad} f_n(w_t, z_t))$, where α_t is the step-size, and where $\text{grad} f_n(w_t, z_t)$ is a Riemannian stochastic gradient, which is a tangent vector at $w_t \in \mathcal{M}$. $\text{grad} f_n(w_t, z_t)$ represents an unbiased estimator of the Riemannian full gradient $\text{grad} f(w_t)$, and the expectation of $\text{grad} f_n(w_t, z_t)$ over the choices of z_t is $\text{grad} f(w_t)$, i.e., $\mathbb{E}_{z_t}[\text{grad} f_n(w_t, z_t)] = \text{grad} f(w_t)$. The update moves from w_t in the stochastic direction $-\text{grad} f_n(w_t, z_t)$ with a step-size α_t while remaining on the manifold \mathcal{M} . This mapping, denoted as $R_w : T_w\mathcal{M} \rightarrow \mathcal{M} : \zeta_w \mapsto R_w(\zeta_w)$, is called *retraction* at w , which maps the tangent bundle $T_w\mathcal{M}$ onto \mathcal{M} with a local rigidity condition that preserves gradients at w . *Exponential mapping* Exp is an instance of the retraction.

Riemannian stochastic variance reduced gradient (R-SVRG): R-SVRG has double loops where a k -th outer loop, called *epoch*, has m_k inner iterations. R-SVRG keeps $\tilde{w}^k \in \mathcal{M}$ after m_{k-1} inner iterations of $(k-1)$ -th epoch, and computes the full Riemannian gradient $\text{grad} f(\tilde{w}^k)$ only for this stored \tilde{w}^k . The algorithm also computes the Riemannian stochastic gradient $\text{grad} f_{i_t^k}(\tilde{w}^k)$ that corresponds to each i_t^k -th sample. Then, picking i_t^k -th sample for each t -th inner iteration of k -th epoch at w_t^k , we calculate ξ_t^k , i.e., by modifying the

Algorithm 1 Algorithm for Riemannian stochastic quasi-Newton with gradient variance reduction (R-SQN-VR).

Require: Update frequency $m_k > 0$, step-size $\alpha_t^k > 0$, memory size L , and cautious update threshold ϵ .

- 1: Initialize \tilde{w}^0 , and calculate the Riemannian full gradient $\text{grad}f(\tilde{w}^0)$.
 - 2: **for** $k = 0, \dots$ **do**
 - 3: Store $w_0^k = \tilde{w}^k$.
 - 4: **for** $t = 0, 1, 2, \dots, m_k - 1$ **do**
 - 5: Choose $i_t^k \in \{1, \dots, N\}$ uniformly at random.
 - 6: Calculate the tangent vector $\tilde{\eta}_t^k$ from \tilde{w}^k to w_t^k by $\tilde{\eta}_t^k = R_{\tilde{w}^k}^{-1}(w_t^k)$.
 - 7: **if** $k > 1$ **then**
 - 8: Transport the stochastic gradient $\text{grad}f_{i_t^k}(w_t^k)$ to $T_{\tilde{w}^k}\mathcal{M}$ by $(\mathcal{T}_{\tilde{\eta}_t^k})^{-1}\text{grad}f_{i_t^k}(w_t^k)$.
 - 9: Calculate $\tilde{\xi}_t^k$ as $\tilde{\xi}_t^k = (\mathcal{T}_{\tilde{\eta}_t^k})^{-1}\text{grad}f_{i_t^k}(w_t^k) - (\text{grad}f_{i_t^k}(\tilde{w}^k) - \text{grad}f(\tilde{w}^k))$.
 - 10: Calculate $\tilde{\mathcal{H}}_t^k \tilde{\xi}_t^k$, transport $\tilde{\mathcal{H}}_t^k \tilde{\xi}_t^k$ back to $T_{w_t^k}\mathcal{M}$ by $\mathcal{T}_{\tilde{\eta}_t^k} \tilde{\mathcal{H}}_t^k \tilde{\xi}_t^k$, and obtain $\mathcal{H}_t^k \xi_t^k$.
 - 11: Update w_{t+1}^k from w_t^k as $w_{t+1}^k = R_{w_t^k}(-\alpha_t^k \mathcal{H}_t^k \xi_t^k)$.
 - 12: **else**
 - 13: Calculate ξ_t^k as $\xi_t^k = \text{grad}f_{i_t^k}(w_t^k) - \mathcal{T}_{\tilde{\eta}_t^k}(\text{grad}f_{i_t^k}(\tilde{w}^k) - \text{grad}f(\tilde{w}^k))$.
 - 14: Update w_{t+1}^k from w_t^k as $w_{t+1}^k = R_{w_t^k}(-\alpha_t^k \xi_t^k)$.
 - 15: **end if**
 - 16: **end for**
 - 17: option I: $\tilde{w}^{k+1} = g_{m_k}(w_1^k, \dots, w_{m_k}^k)$ (or $\tilde{w}^{k+1} = w_t^k$ for randomly chosen $t \in \{1, \dots, m_k\}$).
 - 18: option II: $\tilde{w}^{k+1} = w_{m_k}^k$.
 - 19: Calculate the Riemannian full gradient $\text{grad}f(\tilde{w}^{k+1})$.
 - 20: Calculate the tangent vector η_k from \tilde{w}^k to \tilde{w}^{k+1} by $\eta_k = R_{\tilde{w}^k}^{-1}(\tilde{w}^{k+1})$.
 - 21: Compute $s_k^{k+1} = \mathcal{T}_{\eta_k} \eta_k$, and $y_k^{k+1} = \kappa_k^{-1} \text{grad}f(\tilde{w}^{k+1}) - \mathcal{T}_{\eta_k} \text{grad}f(\tilde{w}^k)$ where $\kappa_k = \|\eta_k\| / \|\mathcal{T}_{R_{\eta_k}} \eta_k\|$.
 - 22: **if** $\langle y_k^{k+1}, s_k^{k+1} \rangle_{\tilde{w}^{k+1}} \geq \epsilon \|s_k^{k+1}\|_{\tilde{w}^{k+1}}^2$ **then**
 - 23: Discard pair (s_{k-L}^k, y_{k-L}^k) when $k > L$, and store pair (s_k^{k+1}, y_k^{k+1}) .
 - 24: **end if**
 - 25: Transport $\{(s_j^k, y_j^k)\}_{j=k-\tau+1}^{k-1} \in T_{\tilde{w}^k}\mathcal{M}$ to $\{(s_j^{k+1}, y_j^{k+1})\}_{j=k-\tau+1}^{k-1} \in T_{\tilde{w}^{k+1}}\mathcal{M}$ by \mathcal{T}_{η_k} , where $\tau = \min(k, L)$.
 - 26: **end for**
-

stochastic gradient $\text{grad}f_{i_t^k}(w_t^k)$ using both $\text{grad}f(\tilde{w}^k)$ and $\text{grad}f_{i_t^k}(\tilde{w}^k)$. Because they belong to different tangent spaces, a simple addition of them is not well-defined because Riemannian manifolds are not vector spaces. Therefore, after $\text{grad}f_{i_t^k}(\tilde{w}^k)$ and $\text{grad}f(\tilde{w}^k)$ are transported to $T_{w_t^k}\mathcal{M}$ by $\mathcal{T}_{\tilde{\eta}_t^k}$, ξ_t^k is set as $\xi_t^k = \text{grad}f_{i_t^k}(w_t^k) - \mathcal{T}_{\tilde{\eta}_t^k}(\text{grad}f_{i_t^k}(\tilde{w}^k) - \text{grad}f(\tilde{w}^k))$, where \mathcal{T} represents *vector transport* from \tilde{w}^k to w_t^k , and $\tilde{\eta}_t^k \in T_{\tilde{w}^k}\mathcal{M}$ satisfies $R_{\tilde{w}^k}(\tilde{\eta}_t^k) = w_t^k$. The vector transport $\mathcal{T} : T\mathcal{M} \oplus T\mathcal{M} \rightarrow T\mathcal{M}, (\eta_w, \xi_w) \mapsto \mathcal{T}_{\eta_w}\xi_w$ is associated with retraction R and all $\xi_w, \zeta_w \in \mathcal{T}_w\mathcal{M}$. It holds that (i) $\mathcal{T}_{\eta_w}\xi_w \in \mathcal{T}_{R(\eta_w)}\mathcal{M}$, (ii) $\mathcal{T}_{0_w}\xi_w = \xi_w$, and (iii) \mathcal{T}_{η_w} is a linear map. *Parallel translation* P is an instance of the vector transport. Consequently, the final update is defined as $w_{t+1}^k = R_{w_t^k}(-\alpha_t^k \xi_t^k)$.

2.2 Proposed R-SQN-VR

We propose a Riemannian stochastic QN method accompanied with a VR method (R-SQN-VR). A straightforward extension is to update the modified stochastic gradient ξ_t^k by premultiplying a linear *inverse Hessian approximation operator* \mathcal{H}_t^k at w_t^k as

$$w_{t+1}^k = R_{w_t^k}(-\alpha_t^k \mathcal{H}_t^k \xi_t^k), \quad (2)$$

where $\mathcal{H}_t^k := \mathcal{T}_{\tilde{\eta}_t^k} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}_{\tilde{\eta}_t^k})^{-1}$ by denoting the inverse Hessian approximation at \tilde{w}^k simply as $\tilde{\mathcal{H}}^k$. Here, \mathcal{T} is an isometric vector transport explained in Section 3. \mathcal{H}_t^k should be positive definite, i.e., $\mathcal{H}_t^k \succ 0$ and is close to the Hessian of f , i.e., $\text{Hess}f(w_t^k)$. It is noteworthy that $\tilde{\mathcal{H}}^k$ is calculated only every outer epoch, and remains to be used for \mathcal{H}_t^k throughout the corresponding k -th epoch.

Curvature pair (s_k^{k+1}, y_k^{k+1}) : This paper particularly addresses the operator $\tilde{\mathcal{H}}^k$ used in L-BFGS intended for a large-scale data. Thus, let s_k^{k+1} and y_k^{k+1} be the variable variation and the gradient variation at $T_{\tilde{w}^{k+1}}\mathcal{M}$, respectively, where the superscript expresses explicitly that they belong to $T_{\tilde{w}^{k+1}}\mathcal{M}$. It should be noted that the curvature pair (s_k^{k+1}, y_k^{k+1}) is calculated at the new $T_{\tilde{w}^{k+1}}\mathcal{M}$ just after k -th epoch finished. Furthermore, after the epoch index k is incremented, the curvature pair must be used only at $T_{\tilde{w}^k}\mathcal{M}$ because the calculation of $\tilde{\mathcal{H}}^k$ is performed only at $T_{\tilde{w}^k}\mathcal{M}$.

The variable variation s_k^{k+1} is calculated from the difference between \tilde{w}^{k+1} and \tilde{w}^k . This is represented by the tangent vector η_k from \tilde{w}^k to \tilde{w}^{k+1} , which is calculated using the inverse of the retraction $R_{\tilde{w}^k}^{-1}(\tilde{w}^{k+1})$. Since η_k belongs to the $T_{\tilde{w}^k}\mathcal{M}$, transporting this onto $T_{\tilde{w}^{k+1}}\mathcal{M}$ yields

$$s_k^{k+1} = \mathcal{T}_{\eta_k}\eta_k (= \mathcal{T}_{\eta_k}R_{\tilde{w}^k}^{-1}(\tilde{w}^{k+1})). \quad (3)$$

The gradient variation y_k^{k+1} is calculated from the difference between the newly calculated full gradient $\text{grad}f(\tilde{w}^{k+1}) \in T_{\tilde{w}^{k+1}}\mathcal{M}$ and the previous one, but transported $\mathcal{T}_{\eta_k}\text{grad}f(\tilde{w}^k) \in T_{\tilde{w}^k}\mathcal{M}$ [20] as

$$y_k^{k+1} = \kappa_k^{-1} \text{grad}f(\tilde{w}^{k+1}) - \mathcal{T}_{\eta_k}\text{grad}f(\tilde{w}^k), \quad (4)$$

where $\kappa_k > 0$ is explained in Section 3.

Inverse Hessian approximation operator $\tilde{\mathcal{H}}^k$: $\tilde{\mathcal{H}}^k$ is calculated using the past curvature pairs. More specifically, $\tilde{\mathcal{H}}^k$ is updated as $\tilde{\mathcal{H}}^{k+1} = (\check{\mathcal{V}}^k)^b \check{\mathcal{H}}_k \check{\mathcal{V}}^k + \rho_k s_k s_k^b$, where $\check{\mathcal{H}}_k = \mathcal{T}_{\eta_k} \circ \tilde{\mathcal{H}}^k \circ \mathcal{T}_{\eta_k}^{-1}$, $\rho_k = 1/\langle y_k, s_k \rangle$, $\check{\mathcal{V}}^k = \text{id} - \rho_k y_k s_k^b$ with identity mapping id [20]. Therein, a^b

denotes the flat of $a \in T_w \mathcal{M}$, i.e., $a^\flat : T_w \mathcal{M} \rightarrow \mathbb{R} : v \rightarrow \langle a, v \rangle_w$. Thus, $\tilde{\mathcal{H}}^k$ depends on $\tilde{\mathcal{H}}^{k-1}$ and (s_{k-1}, y_{k-1}) , and similarly $\tilde{\mathcal{H}}^{k-1}$ depends on $\tilde{\mathcal{H}}^{k-2}$ and (s_{k-2}, y_{k-2}) . Proceeding recursively, $\tilde{\mathcal{H}}^k$ is a function of the initial $\tilde{\mathcal{H}}^0$ and all previous k curvature pairs $\{(s_j, y_j)\}_{j=0}^{k-1}$. Meanwhile, L-BFGS restricts use to the most recent L pairs $\{(s_j, y_j)\}_{j=k-L}^{k-1}$ since (s_j, y_j) with $j < k-L$ are likely to have little curvature information. Based on this idea, L-BFGS performs L updates by the initial $\tilde{\mathcal{H}}^0$. We use the k pairs $\{(s_j, y_j)\}_{j=0}^{k-1}$ when $k < L$.

Now, we consider the final calculation of $\tilde{\mathcal{H}}^k$ used for \mathcal{H}_t^k in the inner iterations (2) of k -th outer epoch using the L most recent curvature pairs. Here, since this calculation is executed at $T_{\tilde{w}^k} \mathcal{M}$ and a Riemannian manifold is in general not a vector space, all the L curvature pairs must be located at $T_{\tilde{w}^k} \mathcal{M}$. To this end, just after the curvature pair is calculated in (3) and (4), the past $(L-1)$ pairs of $\{(s_j^k, y_j^k)\}_{j=k-L+1}^{k-1} \in T_{\tilde{w}^k} \mathcal{M}$ are transported into $T_{\tilde{w}^{k+1}} \mathcal{M}$ by the same vector transport \mathcal{T}_{η_k} used when calculating s_k^{k+1} and y_k^{k+1} . It should be emphasized that this transport is necessary only for every outer epoch instead of every inner loop, and results in drastic reduction of computational complexity in comparison with the straightforward extension of the Euclidean stochastic L-BFGS [24] into the manifold setting. Consequently, the update is defined as

$$\begin{aligned} \tilde{\mathcal{H}}^k &= ((\check{\mathcal{V}}_{k-1}^k)^\flat \cdots (\check{\mathcal{V}}_{k-L}^k)^\flat) \tilde{\mathcal{H}}_0^k (\check{\mathcal{V}}_{k-L}^k \cdots \check{\mathcal{V}}_{k-1}^k) + \cdots \\ &\quad + \rho_{k-2} (\check{\mathcal{V}}_{k-1}^k)^\flat s_{k-2}^k (s_{k-2}^k)^\flat (\check{\mathcal{V}}_{k-1}^k) \\ &\quad + \rho_{k-1} s_{k-1}^k (s_{k-1}^k)^\flat, \end{aligned} \quad (5)$$

where $\check{\mathcal{V}}_j^k = \text{id} - \rho_j y_j^k (s_j^k)^\flat$, and $\tilde{\mathcal{H}}_0^k$ is the initial inverse Hessian approximation. Because $\tilde{\mathcal{H}}_0^k$ is not necessarily $\tilde{\mathcal{H}}^{k-L}$, and because it is any positive definite self-adjoint operator, we use $\tilde{\mathcal{H}}_0^k = \langle s_{k-1}^k, y_{k-1}^k \rangle_{\tilde{w}^k} / \langle y_{k-1}^k, y_{k-1}^k \rangle_{\tilde{w}^k} \text{id}$ similar to the Euclidean case. The practical update of $\tilde{\mathcal{H}}^k$ uses *two-loop recursion* algorithm [26, Section 7.2] in Algorithm A.1 of the supplementary material.

Cautious update: Euclidean L-BFGS fails on non-convex problems because the Hessian approximation has eigenvalues that are away from zero and are not uniformly bounded above. To circumvent this issue, *cautious update* has been proposed in the Euclidean space [27]. By following this, we skip the update of the curvature pair when the following condition is not satisfied;

$$\langle y_k^{k+1}, s_k^{k+1} \rangle_{\tilde{w}^{k+1}} \geq \epsilon \|s_k^{k+1}\|_{\tilde{w}^{k+1}}^2, \quad (6)$$

where $\epsilon > 0$ is a predefined constant parameter. According to this update, the positive definiteness of $\tilde{\mathcal{H}}^k$ is guaranteed as far as $\tilde{\mathcal{H}}^{k-1}$ is positive definite as seen in the proof of Proposition 3.1 in the supplementary material.

Second-order modified stochastic gradient $\mathcal{H}_t^k \xi_t^k$: R-SVRG transports the full gradient $\text{grad} f(\tilde{w}^k)$ and the stochastic gradient $\text{grad} f_{i_t^k}(\tilde{w}^k)$ at $T_{\tilde{w}^k} \mathcal{M}$ into $T_{w_t^k} \mathcal{M}$ to add them to the stochastic gradient $\text{grad} f_{i_t^k}(w_t^k)$ at $T_{w_t^k} \mathcal{M}$. If we follow the same strategy, we must also transport L pairs of $\{(s_j^k, y_j^k)\}_{j=k-L}^{k-1} \in T_{\tilde{w}^k} \mathcal{M}$ into the current $T_{w_t^k} \mathcal{M}$ at every inner iteration. Addressing this problem and the fact that both the full gradient and the curvature pairs belong to the same tangent space $T_{\tilde{w}^k} \mathcal{M}$, we transport $\text{grad} f_{i_t^k}(w_t^k)$ from $T_{w_t^k} \mathcal{M}$ into $T_{\tilde{w}^k} \mathcal{M}$, and complete all the calculations on $T_{\tilde{w}^k} \mathcal{M}$. More specifically, after transporting $\text{grad} f_{i_t^k}(w_t^k)$ as $(\mathcal{T}_{\tilde{\eta}_t^k})^{-1} \text{grad} f_{i_t^k}(w_t^k)$ from w_t^k to \tilde{w}^k using $\tilde{\eta}_t^k (= R_{\tilde{w}^k}^{-1}(w_t^k))$, the modified stochastic gradient

$\tilde{\xi}_t^k \in T_{\tilde{w}^k} \mathcal{M}$ is computed as

$$\tilde{\xi}_t^k = (\mathcal{T}_{\tilde{\eta}_t^k})^{-1} \text{grad} f_{i_t^k}(w_t^k) - (\text{grad} f_{i_t^k}(\tilde{w}^k) - \text{grad} f(\tilde{w}^k)). \quad (7)$$

After calculating $\tilde{\mathcal{H}}_t^k \tilde{\xi}_t^k \in T_{\tilde{w}^k} \mathcal{M}$ using the two-loop recursion algorithm, we obtain $\mathcal{H}_t^k \xi_t^k \in T_{w_t^k} \mathcal{M}$ by transporting $\tilde{\mathcal{H}}_t^k \tilde{\xi}_t^k$ back to $T_{w_t^k} \mathcal{M}$ by $\mathcal{T}_{\tilde{\eta}_t^k} \tilde{\mathcal{H}}_t^k \tilde{\xi}_t^k$. Finally, we update w_{t+1}^k from w_t^k as (2).

3 Convergence analysis

In this section, after defining necessary assumptions, we bound the eigenvalue of \mathcal{H}_t^k . Then, we present a global convergence analysis and a local convergence rate analysis. The concrete proofs are given in the supplementary file.

Assumption 1. We assume below [20];

1. The objective function f and its components f_1, \dots, f_N are twice continuously differentiable.
2. For a sequence $\{w_t^k\}$ generated by Algorithm 1, there exists a compact and connected set $K \subset \mathcal{M}$ such that $w_t^k \in K$ for all $k, t \geq 0$. Also, for each $k \geq 1$, there exists a totally retractive neighborhood Θ_k of \tilde{w}^k such that w_t^k stays in Θ_k for any $t \geq 0$, where the ρ -totally neighborhood Θ of w is a set such that for all $z \in \Theta$, $\Theta \subset R_z(\mathbb{B}(0_z, \rho))$, and $R_z(\cdot)$ is a diffeomorphism on $\mathbb{B}(0_z, \rho)$, which is the ball in $T_w \mathcal{M}$ with center 0_z and radius ρ , where 0_z is the zero vector in $T_z \mathcal{M}$. Furthermore, suppose that there exists $I > 0$ such that $\inf_{k \geq 1} \{\sup_{z \in \Theta_k} \|R_{\tilde{w}^k}^{-1}(z)\|_{\tilde{w}^k}\} \geq I$.
3. The objective function f is strongly retraction-convex with respect to R in Θ . Here, f is said to be strongly retraction-convex in Θ if $f(R_w(t\eta_w))$ for all $w \in \mathcal{M}$ and $\eta_w \in T_w \mathcal{M}$ is strongly convex, i.e., there exist two constants $0 < \lambda < \Lambda$ such that

$$\lambda \leq \frac{d^2 f(R_w(t\eta_w))}{dt^2} \leq \Lambda, \quad (8)$$

for all $w \in \Theta$, all $\|\eta_w\|_w = 1$, and all t such that $R_w(\tau\eta_w) \in \Theta$ for all $\tau \in [0, t]$.

4. The vector transport \mathcal{T} is isometric on \mathcal{M} . It satisfies $\langle \mathcal{T}_{\xi_w} \eta_w, \mathcal{T}_{\xi_w} \zeta_w \rangle_{R_w(\xi_w)} = \langle \eta_w, \zeta_w \rangle_w$ for any $w \in \mathcal{M}$ and $\xi_w, \eta_w, \zeta_w \in T_w \mathcal{M}$.
5. There exists a constant c_0 such that the vector transport \mathcal{T} satisfies the following conditions for all $w, z \in \mathcal{U}$, which is a neighborhood of \tilde{w} :

$$\|\mathcal{T}_{\eta_w} - \mathcal{T}_{R_{\eta_w}}\| \leq c_0 \|\eta_w\|_w, \quad \|\mathcal{T}_{\eta_w}^{-1} - \mathcal{T}_{R_{\eta_w}}^{-1}\| \leq c_0 \|\eta_w\|_w,$$

where \mathcal{T}_R denotes the differentiated retraction, i.e., $\mathcal{T}_{R_{\eta_w}} \xi_w = D R_w(\eta_w)[\xi_w]$, and where $\eta_w, \xi_w \in T_w \mathcal{M}$, and $\eta_w = R_w^{-1}(z)$.

6. The vector transport \mathcal{T} satisfies the locking condition, which is defined as

$$\mathcal{T}_{\eta_w} \xi_w = \kappa \mathcal{T}_{R_{\eta_w}} \xi_w, \quad \text{where } \kappa = \frac{\|\xi_w\|_w}{\|\mathcal{T}_{R_{\eta_w}} \xi_w\|_{R_w(\eta_w)}}, \quad (9)$$

for all $\eta_w, \xi_w \in T_w \mathcal{M}$ and all $w \in \mathcal{M}$.

The following proposition bounds the eigenvalues of \mathcal{H}_t^k .

Proposition 3.1 (Eigenvalue bounds of \mathcal{H}_t^k). *Consider the operator $\tilde{\mathcal{H}}^k := \mathcal{T}_{\eta_t^k} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}_{\eta_t^k})^{-1}$, where $\tilde{\mathcal{H}}^k$ is defined by the recursion in (5). Define the constant $0 < \gamma < \Gamma < \infty$. If Assumption 1 holds, the eigenvalues of \mathcal{H}_t^k is bounded by γ and Γ for all $k \geq 1, t \geq 1$, i.e.,*

$$\gamma \text{id} \preceq \mathcal{H}_t^k \preceq \Gamma \text{id}. \quad (10)$$

It should be noted that, although $-\xi_t^k$ is not generally guaranteed as a descent direction, $\mathbb{E}_{i_t^k}[-\xi_t^k] = -\text{grad}f(w_t^k)$ is a descent direction. Furthermore, the positive definiteness of \mathcal{H}_t^k yields that $-\mathcal{H}_t^k \xi_t^k$ is an average descent direction due to $\mathbb{E}_{i_t^k}[-\mathcal{H}_t^k \xi_t^k] = -\mathcal{H}_t^k \text{grad}f(w_t^k)$.

3.1 Global convergence analysis

This section analyses a global convergence to a critical point starting from any initialization point, which is common in a non-convex setting. We first state the additional assumptions particularly required for this analysis.

Assumption 2. *We assume the following assumptions;*

1. *f is bounded below by a scalar f_{\inf} .*
2. *Since Θ is compact, all continuous functions on Θ can be bounded. Therefore, there exists $S > 0$ such that for all $w \in \Theta$ and $n \in N$, we have*

$$\|\text{grad}f(w)\|_w \leq S \quad \text{and} \quad \|\text{grad}f_n(w)\|_w \leq S. \quad (11)$$

3. *The step-size sequence $\{\alpha_t^k\}$ satisfies*

$$\sum \alpha_t^k = \infty \quad \text{and} \quad \sum (\alpha_t^k)^2 < \infty. \quad (12)$$

Now, we present the global convergence result below.

Theorem 3.2. *Consider Algorithm 1 and suppose Assumptions 1 and 2, and that the mapping $w \mapsto \|\text{grad}f(w)\|_w^2$ has the positive real number that the largest eigenvalue of its Riemannian Hessian is bounded by for all $w \in \mathcal{M}$. Then, we have $\lim_{k \rightarrow \infty} \mathbb{E}[\|\text{grad}f(w_t^k)\|_{w_t^k}^2] = 0$.*

3.2 Local convergence rate analysis

The analysis below presents a convergence rate in neighborhood of a local minimum. This setting is also very common and standard in manifold optimization. For this purpose, we first briefly summarize essential inequalities. They are detailed in the supplementary material.

For all $w, z \in \mathcal{U}$, which is a neighborhood of \bar{w} , the difference between the parallel translation and the vector transport is given with a constant θ as (Lemma D.2)

$$\|\mathcal{T}_\eta \xi - P_\eta \xi\|_z \leq \theta \|\xi\|_w \|\eta\|_w, \quad (13)$$

where $\xi, \eta \in T_w \mathcal{M}$ and $R_w(\eta) = z$. Similarly, as for the difference between the exponential mapping and the retraction, there exist $\tau_1 > 0, \tau_2 > 0$ for all $w \in \mathcal{U}$ and all small length of $\xi \in T_w \mathcal{M}$ such that (Lemma D.3)

$$\tau_1 \text{dist}(w, R_w(\xi)) \leq \|\xi\|_w \leq \tau_2 \text{dist}(w, R_w(\xi)). \quad (14)$$

Then, the variance of ξ_t^k is upper bounded by (Lemma D.5)

$$\mathbb{E}_{i_t^k}[\|\xi_t^k\|_{w_t^k}^2] \leq 4(\beta^2 + \tau_2^2 C^2 \theta^2)(7(\text{dist}(w_t^k, w^*))^2 + 4(\text{dist}(\tilde{w}^k, w^*))^2), \quad (15)$$

where C is the upper bound of $\|\text{grad}f_n(w)\|_w$ for $w \in \Theta$, β is a Lipschitz constant, and θ is the constant in (13).

Now, we are ready to give a local convergence rate as;

Theorem 3.3. *Let \mathcal{M} be a Riemannian manifold and $w^* \in \mathcal{M}$ be a non-degenerate local minimizer of f (i.e., $\text{grad}f(w^*) = 0$ and the Hessian $\text{Hess}f(w^*)$ of f at w^* is positive definite). Suppose Assumption 1 holds. λ and Λ are constants in (8). γ and Γ are constants in Proposition 3.1. Let the constants θ in (13), τ_1 and τ_2 in (14), and β , and C in (15). Let α be a positive number satisfying $\lambda\tau_1^2 > 2\alpha(\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2 C^2 \theta^2))$ and $\gamma\lambda^2\tau_1^2 > 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2 C^2 \theta^2)$. It then follows that for any sequence $\{\tilde{w}^k\}$ generated by Algorithm 1 with a fixed step-size $\alpha_t^k := \alpha$ and $m_k := m$ converging to w^* , there exists $K > 0$ such that for all $k > K$,*

$$\mathbb{E}[(\text{dist}(\tilde{w}^{k+1}, w^*))^2] \leq \frac{2(\Lambda\tau_2^2 + 16m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2 C^2 \theta^2))}{m\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2 C^2 \theta^2))} \mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2]. \quad (16)$$

It should also be noted that, the parallel translation P as the vector transport can lead to a smaller Γ and a larger γ in (10). As a result, this produces a smaller coefficient above, and can result in a faster local convergence rate.

4 Numerical comparisons

This section compares the performance of R-SQN-VR with R-SGD and R-SVRG. While R-SQN-VR and R-SVRG use a fixed step-size $\alpha_t^k = \alpha_0$, R-SGD uses decaying step-size sequences with $\alpha_k = \alpha_0(1 + \alpha_0\nu\lfloor k/m_k \rfloor)^{-1}$, where $\lfloor \cdot \rfloor$ denotes the floor function. We select multiple choices of α_0 and $\nu = 10^{-3}$. We also compare them with R-SD, which is the steepest descent algorithm on Riemannian manifolds with backtracking line search [23, Chapters 4]. All experiments herein use $m_k = 5N$ recommended by [7]. It should be noted that all results except R-SD are the best-tuned results. All experiments are executed in Matlab on a 3.0 GHz Intel Core i7 PC with 16 GB RAM. This paper addresses two problems; the Karcher mean computation problem of symmetric positive-definite (SPD) matrices, and the low-rank matrix completion (MC) problem on the Grassmann manifold. In these problems, full gradient methods become prohibitively computationally expensive when N is very large; the stochastic gradient approach is one promising way to achieve scalability.

4.1 SPD manifold and Karcher mean problem

SPD manifold \mathcal{S}_{++}^d . Let \mathcal{S}_{++}^d be the manifold of $d \times d$ SPD matrices. If we endow \mathcal{S}_{++}^d with the Riemannian metric defined by $\langle \xi_{\mathbf{X}}, \eta_{\mathbf{X}} \rangle_{\mathbf{X}} = \text{trace}(\xi_{\mathbf{X}} \mathbf{X}^{-1} \eta_{\mathbf{X}} \mathbf{X}^{-1})$ at $\mathbf{X} \in \mathcal{S}_{++}^d$, the SPD manifold \mathcal{S}_{++}^d becomes a Riemannian manifold. The explicit formula for the exponential mapping is given by $\text{Exp}_{\mathbf{X}}(\xi_{\mathbf{X}}) = \mathbf{X}^{1/2} \exp(\mathbf{X}^{-1/2} \xi_{\mathbf{X}} \mathbf{X}^{-1/2}) \mathbf{X}^{1/2}$ for any $\xi_{\mathbf{X}} \in T_{\mathbf{X}} \mathcal{S}_{++}^d$ and $\mathbf{X} \in \mathcal{S}_{++}^d$. On the other hand, $R_{\mathbf{X}}(\xi_{\mathbf{X}}) = \mathbf{X} + \xi_{\mathbf{X}} + \frac{1}{2} \xi_{\mathbf{X}} \mathbf{X}^{-1} \xi_{\mathbf{X}}$ proposed in [28] is a retraction, which is symmetric positive-definite for all $\xi_{\mathbf{X}} \in T_{\mathbf{X}} \mathcal{S}_{++}^d$ and $\mathbf{X} \in \mathcal{S}_{++}^d$. The parallel translation on \mathcal{S}_{++}^d along $\eta_{\mathbf{X}}$ is given by $P_{\eta_{\mathbf{X}}}(\xi_{\mathbf{X}}) = \mathbf{X}^{1/2} \mathbf{Y} \mathbf{X}^{-1/2} \xi_{\mathbf{X}} \mathbf{X}^{-1/2} \mathbf{Y} \mathbf{X}^{1/2}$, where

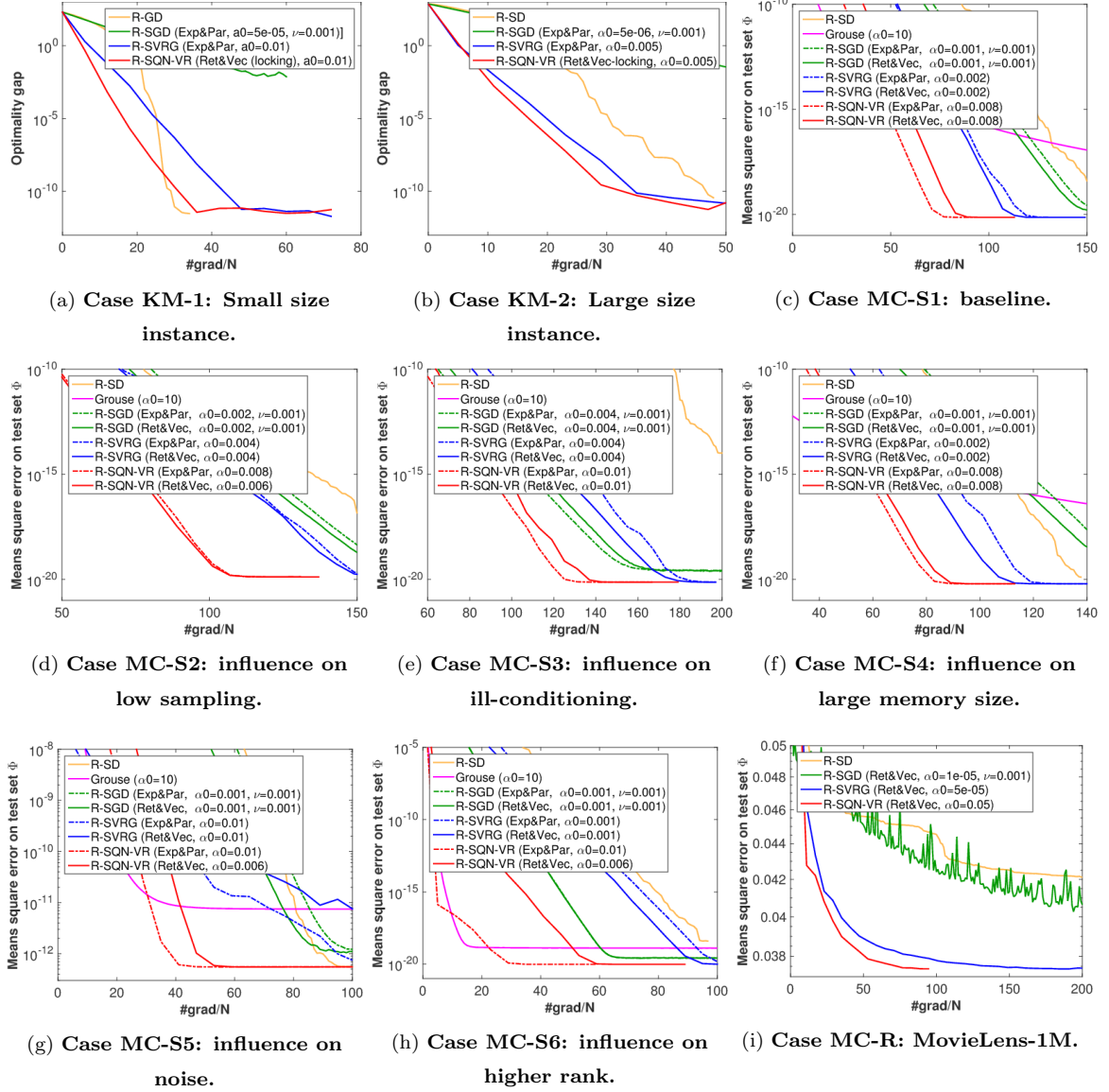


Figure 1: Performance evaluations on Karcher mean problem (KM) low-rank matrix completion problem (MC).

$\mathbf{Y} = \exp(\mathbf{X}^{-1/2} \eta \mathbf{X}^{-1/2} / 2)$. A more efficient algorithm that constructs an isometric vector transport is proposed based on a field of orthonormal tangent bases [29] while satisfying the locking condition (9). We use it in this experiment, and the details are in Appendix E. The logarithm map of \mathbf{Y} at \mathbf{X} is given by $\text{Log}_{\mathbf{X}}(\mathbf{Y}) = \mathbf{X}^{1/2} \log(\mathbf{X}^{-1/2} \mathbf{Y} \mathbf{X}^{-1/2}) \mathbf{X}^{1/2} = \log(\mathbf{Y} \mathbf{X}^{-1}) \mathbf{X}$.

Karcher mean problem on \mathcal{S}_{++}^d . The Karcher mean is introduced as a notion of *mean* on Riemannian manifolds by Karcher [30]. It generalizes the notion of an “average” on a manifold. Given N points on \mathcal{S}_{++}^d with matrix representations $\mathbf{Q}_1, \dots, \mathbf{Q}_N$, the Karcher mean is defined as the solution to the problem

$$\min_{\mathbf{X} \in \mathcal{S}_{++}^d} \frac{1}{2N} \sum_{n=1}^N (\text{dist}(\mathbf{X}, \mathbf{Q}_n))^2, \quad (17)$$

where $\text{dist}(p, q) = \|\log(p^{-1/2} q p^{-1/2})\|_F$ represents the distance along the corresponding geodesic between the elements on \mathcal{S}_{++}^d with respect to the affine-invariant metric. The gradient of the loss function in (17) is computed as $\frac{1}{N} \sum_{n=1}^N -\log(\mathbf{Q}_n \mathbf{X}^{-1}) \mathbf{X}$. The Karcher mean on \mathcal{S}_{++}^d is frequently used for computer vision problems, such as visual object categorization and pose categorization [31]. Since recursive calculations are needed with each visual image, stochastic gradient algorithms become an appealing choice for large datasets.

All experiments use the batch size fixed to 1 and L is 2, and are initialized randomly and are stopped when the number of iterations reaches 60 for R-SD and R-SGD, and 12 for R-SQN-VR and R-SVRG, or the gradient norm gets below 10^{-11} . Figures 1(a) and (b) show the results of the optimality gap in the cases of the small size instance with $N = 1000$ and $d = 3$ (**Case KM-1**) and the large size instance with $N = 10000$ and $d = 10$ (**Case KM-2**), respectively. These results reveal that R-SQN-VR outperforms R-SGD, R-SVRG and R-SD at the convergence speed.

4.2 Grassmann manifold and MC problem

Grassmann manifold $\text{Gr}(r, d)$. A point on the Grassmann manifold is an equivalence class represented by a $d \times r$ orthogonal matrix \mathbf{U} with orthonormal columns, i.e., $\mathbf{U}^T \mathbf{U} = \mathbf{I}$. Two orthogonal matrices express the same element on the Grassmann manifold if they are related by right multiplication of an $r \times r$ orthogonal matrix $\mathbf{O} \in \mathcal{O}(r)$. Equivalently, an element of $\text{Gr}(r, d)$ is identified with a set of $d \times r$ orthogonal matrices $[\mathbf{U}] := \{\mathbf{U}\mathbf{O} : \mathbf{O} \in \mathcal{O}(r)\}$. That is, $\text{Gr}(r, d) := \text{St}(r, d) / \mathcal{O}(r)$, where $\text{St}(r, d)$ is the *Stiefel manifold* that is the set of matrices of size $d \times r$ with orthonormal columns. The Grassmann manifold has the structure of a Riemannian quotient manifold [23, Section 3.4].

The exponential mapping for the Grassmann manifold from $\mathbf{U}(0) := \mathbf{U} \in \text{Gr}(r, d)$ in the direction of $\xi \in T_{\mathbf{U}(0)} \text{Gr}(r, d)$ is given in a closed form as [23, Section 5.4] $\mathbf{U}(t) = [\mathbf{U}(0) \mathbf{V} \quad \mathbf{W}] [\cos t\Sigma; \sin t\Sigma] \mathbf{V}^T$, where $\xi = \mathbf{W} \Sigma \mathbf{V}^T$ is the singular value decomposition (SVD) of ξ with rank r . The $\sin(\cdot)$ and $\cos(\cdot)$ operations are performed only on the diagonal entries. The parallel translation of $\zeta \in T_{\mathbf{U}(0)} \text{Gr}(r, d)$ on the Grassmann manifold along $\gamma(t)$ with $\dot{\gamma}(0) = \mathbf{W} \Sigma \mathbf{V}^T$ is given in a closed form by $\zeta(t) = ([\mathbf{U}(0) \mathbf{V} \quad \mathbf{W}] [-\sin t\Sigma; \cos t\Sigma] \mathbf{W}^T + (\mathbf{I} - \mathbf{W} \mathbf{W}^T)) \zeta$. The logarithm map of $\mathbf{U}(t)$ at $\mathbf{U}(0)$ on the Grassmann manifold is given by $\xi = \text{Log}_{\mathbf{U}(0)}(\mathbf{U}(t)) = \mathbf{W} \arctan(\Sigma) \mathbf{V}^T$, where $\mathbf{W} \Sigma \mathbf{V}^T$ is the SVD of $(\mathbf{U}(t) - \mathbf{U}(0) \mathbf{U}(0)^T \mathbf{U}(t)) (\mathbf{U}(0)^T \mathbf{U}(t))^{-1}$ with rank r . Furthermore, a popular retraction is $R_{\mathbf{U}(0)}(\xi) = \text{qf}(\mathbf{U}(0) + t\xi) (= \mathbf{U}(t))$ which extracts the orthonormal factor based on QR decomposition, and a popular

vector transport uses an orthogonal projection of $t\xi$ to the horizontal space at $\mathbf{U}(t)$, i.e., $(\mathbf{I} - \mathbf{U}(t)\mathbf{U}(t)^T)t\xi$ [23].

Matrix completion problem. The matrix completion problem is completing an incomplete matrix \mathbf{X} , say of size $d \times N$, from a small number of entries by assuming that the latent structure of the matrix is low-rank. If Ω is the set of known indices in \mathbf{X} , the rank- r matrix completion problem amounts to solving $\min_{\mathbf{U}, \mathbf{A}} \|\mathcal{P}_\Omega(\mathbf{U}\mathbf{A}) - \mathcal{P}_\Omega(\mathbf{X})\|_F^2$, where $\mathbf{U} \in \mathbb{R}^{d \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times N}$, and the operator \mathcal{P}_Ω acts as $\mathcal{P}_\Omega(\mathbf{X}_{ij}) = \mathbf{X}_{ij}$ if $(i, j) \in \Omega$ and $\mathcal{P}_\Omega(\mathbf{X}_{ij}) = 0$ otherwise. Partitioning $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$, the previous problem is equivalent to

$$\min_{\mathbf{U} \in \mathbb{R}^{d \times r}, \mathbf{a}_n \in \mathbb{R}^r} \frac{1}{N} \sum_{n=1}^N \|\mathcal{P}_{\Omega_n}(\mathbf{U}\mathbf{a}_n) - \mathcal{P}_{\Omega_n}(\mathbf{x}_n)\|_2^2, \quad (18)$$

where $\mathbf{x}_n \in \mathbb{R}^d$ and the operator \mathcal{P}_{Ω_n} is the sampling operator for the n -th column. Given \mathbf{U} , \mathbf{a}_n admits a closed form solution. Consequently, the problem only depends on the column space of \mathbf{U} and is on $\text{Gr}(r, d)$ [32].

We first consider a synthetic dataset. The proposed algorithm is also compared with Grouse [1], a state-of-the-art stochastic gradient algorithm on the Grassmann manifold. Algorithms are initialized randomly as suggested in [33]. The multiple choices of α_0 are $\{10^{-3}, 5 \times 10^{-3}, \dots, 10^{-1}, 5 \times 10^{-1}\}$ for R-SGD, R-SVRG and R-SQN-VR, and $\{5 \times 10^{-1}, 1, 5, 10\}$ for Grouse. We set explicitly the condition number, denoted as CN, of the matrix, which represents the ratio of the maximal and the minimal singular values of the matrix. We also set the over-sampling ratio (OS) which prescribes the number of known entries. An OS of 5 means that we randomly and uniformly select known entries of $5(N + d - r)r$ a priori out of the total Nd entries. The Gaussian noise is also added with the noise level σ as suggested in [33]. The batch size is fixed to 10. This experiment evaluates two combinations of vector transport and retraction, namely, the parallel translation and the exponential mapping, and the projection-based vector transport and the QR-decomposition-based retraction. While the former combination satisfies the necessary conditions for the convergence, the latter does not, but is computationally more efficient than the former. The baseline problem instance is the case of the number of samples $N = 5000$, the dimension $d = 500$, the memory size $L = 5$, OS = 6, CN = 5, and $\sigma = 10^{-10}$ (**Case MC-S1**). Additionally, we evaluate the lower-sampling case with OS = 4 (**Case MC-S2**), the ill-conditioning case with CN = 10 (**Case MC-S3**), the larger memory size case with $L = 10$ (**Case MC-S4**), the higher noise case with $\sigma = 10^{-6}$ (**Case MC-S5**), and the higher rank case with $r = 10$ (**Case MC-S6**). The results of the MSE on test set Φ , which is different from the training set Ω , are shown in Figures 1(c)-(h), respectively. This gives the prediction accuracy of missing elements. From the figures, we confirm the superior performance of R-SQN-VR.

Finally, we compare the algorithms on a real-world dataset, the MovieLens-1M dataset¹. It contains a million ratings for 3952 movies (N) of 6040 users (d). We further randomly split this set into 80/10/10 percent data out of the entire data as train/validation/test partitions. α_0 is chosen from $\{10^{-5}, 5 \times 10^{-5}, \dots, 10^{-2}, 5 \times 10^{-2}\}$, the rank r is $\{10, 20\}$, and L is 5. The QR decomposition-based vector transport is used assuming a practical deployment due to the size of datasets. The algorithms are terminated when the tendency is observed that the MSE on the validation set increases or the number of the iteration reaches 100. Figure 1(i) shows the results when $r = 20$ on test sets of all the algorithms except Grouse, which faces issues

¹<http://grouplens.org/datasets/movielens/>

with convergence on this data set (**Case MC-R**). R-SQN-VR shows much faster convergence than others. Table 1 also shows the best result when r is 10 and 20 for each algorithm with the lowest test MSEs when the algorithm stopped for five runs. This also shows that the proposed R-SQN-VR gives faster convergences and lower test MSE than other algorithms, especially when r is larger.

Table 1: Test MSE on Φ and # of gradient (MovieLens-1M).

r	Algorithm	#grad/ N	MSE on Φ
10	R-SG	$399.6 \pm 8.9_{-1}$	$2.597056_{-4} \pm 3.2_{-6}$
	R-SGD	$375.0 \pm 1.8_1$	$2.629046_{-4} \pm 1.0_{-6}$
	R-SVRG	$205.4 \pm 4.1_1$	$2.529259_{-4} \pm 4.7_{-7}$
	R-SQN-VR	$116.6 \pm 3.6_1$	$2.523758_{-4} \pm 3.4_{-7}$
20	R-SG	400 ± 0	$1.950905_{-4} \pm 2.1_{-6}$
	R-SGD	382.0 ± 2.8	$1.975479_{-4} \pm 9.6_{-7}$
	R-SVRG	$356.0 \pm 5.5_1$	$1.896678_{-4} \pm 4.8_{-7}$
	R-SQN-VR	$137.0 \pm 5.1_1$	$1.888096_{-4} \pm 5.0_{-7}$

The subscript k indicates a scale of 10^k .

5 Conclusions

We have proposed a Riemannian stochastic quasi-Newton algorithm with variance reduction (R-SQN-VR). We particularly addressed retraction and vector transport. The proposed algorithm stems from the algorithm in Euclidean space, but is now extended to Riemannian manifolds. The central difficulty of averaging, adding, and subtracting multiple gradients on a Riemannian manifold is handled by exploiting vector transport and retraction. We proved that R-SQN-VR generates globally convergent sequences with a decaying step-size condition and is locally linearly convergent under some natural assumptions. Numerical comparisons suggested the superior performance of R-SQN-VR on various benchmarks.

References

- [1] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Allerton*, pages 704–711, 2010.
- [2] B. Mishra and R. Sepulchre. R3MC: A Riemannian three-factor algorithm for low-rank matrix completion. In *IEEE CDC*, pages 1137–1142, 2014.
- [3] H. Kasai and B. Mishra. Low-rank tensor completion: a Riemannian manifold preconditioning approach. In *ICML*, 2016.
- [4] G. Meyer, S. Bonnabel, and R. Sepulchre. Linear regression under fixed-rank constraints: A Riemannian approach. In *ICML*, 2011.
- [5] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Trans. on Automatic Control*, 58(9):2217–2229, 2013.
- [6] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, pages 400–407, 1951.
- [7] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.
- [8] N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671, 2012.
- [9] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *JMLR*, 14:567–599, 2013.
- [10] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, 2014.
- [11] S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *ICML*, 2016.
- [12] N. N. Schraudolph, J. Yu, and S. Gunter. A stochastic quasi-Newton method for online convex optimization. In *AISTATS*, 2007.
- [13] A. Mokhtari and A. Ribeiro. RES: Regularized stochastic BFGS algorithm. *IEEE Trans. on Signal Process.*, 62(23):6089–6104, 2014.
- [14] X. Wang, S. Ma, D. Goldfarb, and W. Liu. Stochastic quasi-Newton methods for non-convex stochastic optimization. *arXiv preprint arXiv:1607.0123*, 2016.
- [15] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-Newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2), 2016.
- [16] P. Moritz, R. Nishihara, and M. I. Jordan. A linearly-convergent stochastic L-BFGS algorithm. In *AISTATS*, pages 249–258, 2016.
- [17] R. Kolte, M. Erdogdu, and A. Ozgur. Accelerating SVRG via second-order information,. In *OPT2015*, 2015.

- [18] D. Gabay. Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37(2):177–219, 1982.
- [19] W. Ring and B. Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 22(2):596–627, 2012.
- [20] W. Huang, K. A. Gallivan, and P.-A. Absil. A Broyden class of quasi-Newton methods for Riemannian optimization. *SIAM J. Optim.*, 25(3):1660–1685, 2015.
- [21] H. Sato, H. Kasai, and B. Mishra. Riemannian stochastic variance reduced gradient. *arXiv preprint: arXiv:1702.05594*, 2017.
- [22] H. Zhang, S. J. Reddi, and S. Sra. Fast stochastic optimization on Riemannian manifolds. In *NIPS*, 2016.
- [23] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.
- [24] A. Mokhtari and A. Ribeiro. Global convergence of online limited memory BFGS. *JMLR*, 16:3151–3181, 2015.
- [25] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt: a Matlab toolbox for optimization on manifolds. *JMLR*, 15(1):1455–1459, 2014.
- [26] J. Nocedal and Wright S.J. *Numerical Optimization*. Springer, New York, USA, 2006.
- [27] D. Li and M. Fukushima. On the global convergence of BFGS method for nonconvex unconstrained optimization. *SIAM J. Optim.*, 11(4):1054–1064, 2011.
- [28] B. Jeuris, R. Vandebril, and B. Vandereycken. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *ETNA*, 2012.
- [29] X. Yuana, P.-A. Huang, W. Absil, and K. A. Gallivan. A Riemannian limited-memory BFGS algorithm for computing the matrix geometric mean. In *ICCS*, 2016.
- [30] H Karcher. Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.*, 30(5):509–541, 1977.
- [31] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(12), 2015.
- [32] N. Boumal and P.-A. Absil. Low-rank matrix completion via preconditioned optimization on the Grassmann manifold. *Linear Algebra and its Applications*, 475:200–239, 2015.
- [33] D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numer. Math.*, 54(2):447–468, 2014.
- [34] B. Mishra and R. Sepulchre. Riemannian preconditioning. *SIAM J. Optim.*, 635-660, 2016.
- [35] L. Bottou, F. Curtis, and J. Nocedal. Optimization mehtods for large-scale machine learning. *arXiv preprint: arXiv:1606.04838*, 2016.

- [36] W. Huang, P.-A. Absil, and K. A. Gallivan. A Riemannian symmetric rank-one trust-region method. *Math. Program., Ser. A*, 150:179–216, 2015.
- [37] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

A Two-loop Hessian inverse updating algorithm

Algorithm A.1 Hessian Inverse Updating

Require: Pair-updating counter t , memory depth τ , correction pairs $\{s_u^k, y_u^k\}_{u=k-\tau}^{k-1}$, gradient p .

- 1: $p_0 = p$.
- 2: $\mathcal{H}_k^0 = \gamma_k \text{id} = \frac{\langle s_t^k, y_t^k \rangle}{\langle y_t^k, y_t^k \rangle} \text{id}$.
- 3: **for** $u = 0, 1, 2, \dots, \tau - 1$ **do**
- 4: $\rho_{k-u} = 1 / \langle s_{k-u-1}^k, y_{k-u-1}^k \rangle$.
- 5: $\alpha_u = \rho_{k-u-1} \langle s_{k-u-1}^k, p_u \rangle$.
- 6: $p_{u+1} = p_u - \alpha_u y_{k-u-1}^k$.
- 7: **end for**
- 8: $q_0 = \mathcal{H}_k^0 p_\tau$.
- 9: **for** $u = 0, 1, 2, \dots, \tau - 1$ **do**
- 10: $\beta_u = \rho_{k-\tau+u} \langle y_{k-\tau+u}^k, q_u \rangle$.
- 11: $q_{u+1} = q_u + (\alpha_{\tau-u-1} - \beta_u) s_{k-\tau+u}^k$.
- 12: **end for**
- 13: $q = q_\tau$.

B Proof of Proposition 3.1

This section presents the proof of Proposition 3.1, which is an essential proposition that bounds the eigenvalues of \mathcal{H}_t^k at w_t^k , i.e., $\mathcal{H}_t^k := \mathcal{T}_{\tilde{\eta}_t^k} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}_{\tilde{\eta}_t^k})^{-1}$. To this end, we particularly use the Hessian approximation operator $\tilde{\mathcal{B}}^k = (\tilde{\mathcal{H}}^k)^{-1}$ as opposed to $\tilde{\mathcal{H}}^k$. Since mentioned in the algorithm description, we consider the curvature information for $\tilde{\mathcal{H}}^k$ at \tilde{w}^k , i.e., every outer epoch, and reuse this $\tilde{\mathcal{H}}^k$ in the calculation of the second-order modified stochastic gradient $\mathcal{H}_t^k \xi_t^k$ at w_t^k . Thereby, the way of the proof consists of two steps as follows;

1. We first address the bounds of $\tilde{\mathcal{H}}^k$ at \tilde{w}^k . The main task of the proof is to bound the Hessian operator $\tilde{\mathcal{B}}^k = (\tilde{\mathcal{H}}^k)^{-1}$.
2. Next, we bound \mathcal{H}_t^k at w_t^k based on the bounds of $\tilde{\mathcal{H}}^k$ at \tilde{w}^k .

It should be noted that, in this section, the curvature pair $\{s_j^k, y_j^k\}_{j=k-L}^{k-1} \in T_{\tilde{w}^k} \mathcal{M}$ is simply notated as $\{s_j, y_j\}_{j=k-L}^{k-1}$, and we omit the subscript \tilde{w}^k for a Riemannian metric $\langle \cdot, \cdot \rangle_{\tilde{w}^k}$ when the tangent space to be considered is clear.

B.1 Preliminary lemmas

This section first states some essential lemmas. The literature [23] generalizes a Taylor's theorem to Riemannian manifolds. However, it addresses the exponential mapping instead of the retraction. Therefore, [20] applies Taylor's theorem on the retraction by newly introducing a function along a curve on the manifold. Here, we denote $f(R_{w_k}(t\eta_k/\|\eta_k\|))$ for a twice continuously differentiable objective function. Since f is strongly retraction-convex on Θ by

Assumption 1.3, there exist constants $0 < \lambda < \Lambda$ such that $\lambda \leq \frac{d^2 f(R_{w_k}(t\eta_k/\|\eta_k\|))}{dt^2} \leq \Lambda$ for all $t \in [0, \alpha_t^k \|\eta_k\|]$ as in (8). From Taylor's theorem, we obtain

Lemma B.1 (In Lemma 3.2 in [20]). *Under Assumptions 1.1, 1.2 and 1.3, there exists λ such that*

$$f(w_{t+1}^k) - f(w_t^k) \geq \langle \text{grad} f(w_t^k), \alpha_t^k \eta_k \rangle_{w_t^k} + \frac{1}{2} \lambda (\alpha_t^k \|\eta_k\|)^2. \quad (\text{A.1})$$

There also exists Λ such that

$$f(w_{t+1}^k) - f(w_t^k) \leq \langle \text{grad} f(w_t^k), \alpha_t^k \eta_k \rangle_{w_t^k} + \frac{1}{2} \Lambda (\alpha_t^k \|\eta_k\|)^2. \quad (\text{A.2})$$

Proof. From Taylor's theorem, we have

$$\begin{aligned} f(w_{t+1}^k) - f(w_t^k) &= f(R_{w_k}(\alpha_t^k \eta_k)) - f(R_{w_k}(0)) \\ &= \frac{df(R_{w_k}(0))}{dt} \alpha_t^k \|\eta_k\| + \frac{1}{2} \frac{d^2 f(R_{w_k}(p\eta_k/\|\eta_k\|))}{dt^2} (\alpha_t^k \|\eta_k\|)^2 \\ &= \langle \text{grad} f(w_t^k), \alpha_t^k \eta_k \rangle_{w_t^k} + \frac{1}{2} \frac{d^2 f(R_{w_k}(p\eta_k/\|\eta_k\|))}{dt^2} (\alpha_t^k \|\eta_k\|)^2 \\ &\geq \langle \text{grad} f(w_t^k), \alpha_t^k \eta_k \rangle_{w_t^k} + \frac{1}{2} \lambda (\alpha_t^k \|\eta_k\|)^2, \end{aligned}$$

where $0 \leq p \leq \alpha_t^k \|\eta_k\|$, and the inequality uses (8). This yields (A.1). Similarly, (A.2) is also derived. This completes the proof. \square

Lemma B.2 (Lemma 3.3 in [20]). *Under Assumptions 1.1, 1.2, 1.3, 1.4 and 1.6, there exist two constants $0 < \lambda < \Lambda$ such that*

$$\lambda \leq \frac{\langle s_k, y_k \rangle}{\langle s_k, s_k \rangle} \leq \Lambda \quad (\text{A.3})$$

for all k . Constants λ and Λ can be chosen as in (8).

Proof. From Lemma B.1, the proof of this lemma is given, but we omit it. The reader can see the full proof in Lemma 3.3 in [20]. \square

Lemma B.3. *Suppose Assumption 1 holds, there exist a constant $0 < \mu_1$ then for all k such that*

$$\mu_1 \leq \frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle}. \quad (\text{A.4})$$

Proof. This is given by applying Cauchy-Schwarz inequality to the condition (6) recursively. More specifically, (6) yields that

$$\epsilon \|s_k\|^2 \leq \langle y_k, s_k \rangle \leq \|y_k\| \|s_k\|,$$

and considering the most left and right terms, we obtain

$$\|s_k\| \leq \frac{1}{\epsilon} \|y_k\|.$$

Substituting this into the above equation yields

$$\langle s_k, y_k \rangle \leq \|s_k\| \|y_k\| \leq \frac{1}{\epsilon} \|y_k\|^2.$$

Consequently, we obtain

$$\frac{\|y_k\|^2}{\langle s_k, y_k \rangle} \geq \epsilon.$$

This complete the claim by denoting ϵ as μ_1 . \square

Lemma B.4 (Lemma 3.9 in [20]). *Suppose Assumption 1 holds, there exist a constant $0 < \mu_2$ for all k such that*

$$\frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle} \leq \mu_2. \quad (\text{A.5})$$

Proof. (Lemma 3.9 in [20]) The complete proof is given in [20], but this paper provides it for the subsequent analysis. Define $y_k^P = \text{grad}f(\tilde{w}^{k+1}) - P_{\gamma_k}^{1 \leftarrow 0} \text{grad}f(\tilde{w}^k)$, where $\gamma_k(t) = R_{\tilde{w}^k}(t\alpha_k\eta_k)$, i.e., the retraction curve connecting w_k and w_{k+1} , and P_{γ_k} is the parallel translation along $\gamma_k(t)$. We have $\|P_{\gamma_k}^{1 \leftarrow 0} y_k^P - \bar{H}_k \alpha_k \eta_k\| \leq b_0 \|\alpha_k \eta_k\|^2 = b_0 \|s_k\|^2$, where $\bar{H}_k = \int_0^1 P_{\gamma_k}^{0 \leftarrow t} \text{Hess}f(\gamma_k(t)) P_{\gamma_k}^{t \leftarrow 0} dt$ and $b_0 > 0$. It follows that

$$\begin{aligned} \|y_k\| &\leq \|y_k - y_k^P\| + \|y_k^P\| \\ &= \|y_k - y_k^P\| + \|P_{\gamma_k}^{0 \leftarrow 1} y_k^P\| \\ &\leq \|y_k - y_k^P\| + \|P_{\gamma_k}^{0 \leftarrow 1} y_k^P - \bar{H}_k \alpha_k \eta_k\| + \|\bar{H}_k \alpha_k \eta_k\| \\ &\leq \|\text{grad}f(\tilde{w}^{k+1})/\kappa_k - \mathcal{T}_{\alpha_k \eta_k} \text{grad}f(\tilde{w}^k) - \text{grad}f(\tilde{w}^{k+1}) + P_{\gamma_k}^{0 \leftarrow 1} \text{grad}f(\tilde{w}^k)\| \\ &\quad + \|\bar{H}_k \alpha_k \eta_k\| + b_0 \|s_k\|^2 \\ &\leq \|\text{grad}f(\tilde{w}^{k+1})/\kappa_k - \text{grad}f(\tilde{w}^{k+1})\| + \|P_{\gamma_k}^{0 \leftarrow 1} \text{grad}f(\tilde{w}^k) - \mathcal{T}_{\alpha_k \eta_k} \text{grad}f(\tilde{w}^k)\| \\ &\quad + \|\bar{H}_k \alpha_k \eta_k\| + b_0 \|s_k\|^2 \\ &\leq b_1 \|s_k\| \|\text{grad}f(\tilde{w}^{k+1})\| + b_2 \|s_k\| \|\text{grad}f(\tilde{w}^k)\| + b_3 \|s_k\| + b_0 \|s_k\|^2 \\ &\leq b_4 \|s_k\|, \end{aligned} \quad (\text{A.6})$$

where b_1, b_2, b_3 , and $b_4 > 0$. Therefore, by Lemma B.2, we have

$$\frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle} \leq \frac{\langle y_k, y_k \rangle}{\lambda \langle s_k, s_k \rangle} \leq \frac{b_4^2}{\lambda}. \quad (\text{A.7})$$

This complete the proof. \square

Remark B.5. *From the proof of Lemma B.4, if the parallel translation is used for vector transport, i.e., $\mathcal{T} = P$, the first two terms in (A.6) are equal to zero, and the upper bound μ_2 in (A.5) can get smaller than that of the case in the vector transport.*

Now, we attempt to bound $\text{trace}(\hat{\hat{\mathcal{B}}})$ and $\det(\hat{\hat{\mathcal{B}}})$ in order to bound the eigenvalues of $\tilde{\mathcal{H}}^k$, where a *hat* denotes the coordinate expression of the operator. The basic structure of the proof follows stochastic L-BFGS methods in the Euclidean space, e.g., [15, 24]. Nevertheless, some special treatments considering the Riemannian setting and the lemmas earlier are required. It should be noted that $\text{trace}(\hat{\hat{\mathcal{B}}})$ and $\det(\hat{\hat{\mathcal{B}}})$ do not depend on the chosen basis.

Lemma B.6 (Bounds of trace and determinant of $\tilde{\mathcal{B}}^k$). *Consider the recursion of $\tilde{\mathcal{B}}_u^k$ as*

$$\tilde{\mathcal{B}}_{u+1}^k = \tilde{\mathcal{B}}_u^k - \frac{\tilde{\mathcal{B}}_u^k s_{k-\tau+u} (\tilde{\mathcal{B}}_u^k s_{k-\tau+u})^b}{(\tilde{\mathcal{B}}_u^k s_{k-\tau+u})^b s_{k-\tau+u}} + \frac{y_{k-\tau+u} y_{k-\tau+u}^b}{y_{k-\tau+u}^b s_{k-\tau+u}}, \quad (\text{A.8})$$

where $\tilde{\mathcal{B}}_u^k = \mathcal{T}_{\eta_k} \tilde{\mathcal{B}}_u^k (\mathcal{T}_{\eta_k})^{-1}$ for $u = 0, \dots, \tau - 1$. The Hessian approximation at k -th outer epoch is $\tilde{\mathcal{B}}^k = \tilde{\mathcal{B}}_\tau^k$ when $u = \tau - 1$. Then, consider the Hessian approximation $\hat{\mathcal{B}}^k = \tilde{\mathcal{B}}_\tau^k$ in (A.8) with $\tilde{\mathcal{B}}_0^k = \gamma_k^{-1} \text{id}$. If Assumption 1 holds, the $\text{trace}(\hat{\mathcal{B}}^k)$ in a coordinate expression of $\tilde{\mathcal{B}}^k$ is uniformly upper bounded for all $k \geq 1$,

$$\text{trace}(\hat{\mathcal{B}}^k) \leq (M + \tau) \mu_2. \quad (\text{A.9})$$

where M is the dimension of \mathcal{M} . Similarly, if Assumption 1 holds, the $\det(\hat{\mathcal{B}}^k)$ in a coordinate expression of $\tilde{\mathcal{B}}^k$ is uniformly lower bounded for all k ,

$$\det(\hat{\mathcal{B}}^k) \geq \mu_1^M \left[\frac{\lambda}{(M + \tau) \mu_2} \right]^\tau. \quad (\text{A.10})$$

Here, a hat expression represents the coordinate expression of an operator.

Proof. The proof can be completed parallel to the Euclidean case [16]. We use a *hat* symbol in order to represent the coordinate expression of the operator $\tilde{\mathcal{B}}_{u+1}^k$ and $\tilde{\mathcal{B}}_u^k$ in update formula (A.8). Because \mathcal{T} is an isometry vector transport, \mathcal{T}_{η_k} is invertible for all k . Accordingly, $\text{trace}(\hat{\mathcal{B}}^k)$ and $\det(\hat{\mathcal{B}}^k)$ can be reformulated as

$$\text{trace}(\hat{\mathcal{B}}^k) = \text{trace}(\hat{\mathcal{T}}_{\eta_k} \hat{\mathcal{B}}^k \hat{\mathcal{T}}_{\eta_k}^{-1}) = \text{trace}(\hat{\mathcal{B}}^k), \quad (\text{A.11})$$

$$\det(\hat{\mathcal{B}}^k) = \det(\hat{\mathcal{T}}_{\eta_k} \hat{\mathcal{B}}^k \hat{\mathcal{T}}_{\eta_k}^{-1}) = \det(\hat{\mathcal{B}}^k). \quad (\text{A.12})$$

We first consider the trace lower bound of $\text{trace}(\hat{\mathcal{B}}_\tau^k)$ from (A.11) and (A.8) as

$$\begin{aligned} \text{trace}(\hat{\mathcal{B}}_{u+1}^k) &= \text{trace}(\hat{\mathcal{B}}_u^k) - \frac{\langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, \hat{\mathcal{B}}_u^k s_{k-\tau+u} \rangle}{\langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, s_{k-\tau+u} \rangle} + \frac{\langle y_{k-\tau+u}, y_{k-\tau+u} \rangle}{\langle y_{k-\tau+u}, s_{k-\tau+u} \rangle} \\ &= \text{trace}(\hat{\mathcal{B}}_u^k) - \frac{\|\hat{\mathcal{B}}_u^k s_{k-\tau+u}\|^2}{\langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, s_{k-\tau+u} \rangle} + \frac{\|y_{k-\tau+u}\|^2}{\langle y_{k-\tau+u}, s_{k-\tau+u} \rangle}. \end{aligned}$$

Here, the positive definiteness of $\hat{\mathcal{B}}_u^k$ guarantees the negativity of the second term. Therefore, the bound of the third term yields from Lemma B.4 as

$$\text{trace}(\hat{\mathcal{B}}_{u+1}^k) \leq \text{trace}(\hat{\mathcal{B}}_u^k) + \mu_2.$$

By calculating recursively this for $u = 0, \dots, \tau - 1$, we can conclude that

$$\text{trace}(\hat{\mathcal{B}}_u^k) \leq \text{trace}(\hat{\mathcal{B}}_0^k) + u \mu_2.$$

All that is left is to bound $\text{trace}(\hat{\mathcal{B}}_0^k)$. For this purpose, we consider the definition $\hat{\mathcal{B}}_0^k = \text{id}/\gamma_k$ where, as a common choice in L-BFGS in the Euclidean, $\gamma_k = \frac{\langle s_k, y_k \rangle}{\langle y_k, y_k \rangle}$, and we obtain

$$\text{trace}(\hat{\mathcal{B}}_0^k) = \text{trace}\left(\frac{\mathbf{I}}{\gamma_k}\right) = \frac{M}{\gamma_k} = M \frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle} \leq M \mu_2.$$

Consequently, we obtain

$$\text{trace}(\hat{\mathcal{B}}_u^k) \leq (M + u)\mu_2. \quad (\text{A.13})$$

Plugging $u = \tau$ into (A.13) yields the claim (A.9). For $k = 0$, we have $\gamma_k = \gamma_0 = 1$ and (A.13) reduces to $\text{trace}(\hat{\mathcal{B}}_0^k) = M$ while (A.13) results in $\text{trace}(\hat{\mathcal{B}}_u^k) \leq (1 + u)\mu_2$.

We consider the determinant lower bound of $\det(\hat{\mathcal{B}}_{k,\tau}^k)$ from (A.8) as

$$\begin{aligned} \det(\hat{\mathcal{B}}_{u+1}^k) &= \det(\hat{\mathcal{B}}_u^k) \det \left(\mathbf{I} - \frac{s_{k-\tau+u}(\hat{\mathcal{B}}_u^k s_{k-\tau+u})^T}{\langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, s_{k-\tau+u} \rangle} + \frac{(\hat{\mathcal{B}}_u^k)^{-1} y_{k-\tau+u} y_{k-\tau+u}^T}{\langle y_{k-\tau+u}, s_{k-\tau+u} \rangle} \right) \\ &= \det(\hat{\mathcal{B}}_u^k) \det \left(\frac{(\hat{\mathcal{B}}_u^k s_{k-\tau+u})^T}{\langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, s_{k-\tau+u} \rangle} (\hat{\mathcal{B}}_u^k)^{-1} y_{k-\tau+u} \right) \\ &= \det(\hat{\mathcal{B}}_u^k) \frac{\langle s_{k-\tau+u}, y_{k-\tau+u} \rangle}{\langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, s_{k-\tau+u} \rangle} \\ &= \det(\hat{\mathcal{B}}_u^k) \frac{\langle s_{k-\tau+u}, y_{k-\tau+u} \rangle}{\|s_{k-\tau+u}\|^2} \frac{\|s_{k-\tau+u}\|^2}{\langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, s_{k-\tau+u} \rangle} \\ &\geq \det(\hat{\mathcal{B}}_u^k) \frac{\lambda}{\lambda_{\max}(\hat{\mathcal{B}}_u^k)} \\ &\geq \det(\hat{\mathcal{B}}_u^k) \frac{\lambda}{\text{trace}(\hat{\mathcal{B}}_u^k)} \\ &\geq \det(\hat{\mathcal{B}}_u^k) \frac{\lambda}{(M + \tau)\mu_2}. \end{aligned} \quad (\text{A.14})$$

Regarding the second equality, we obtain it from the formula $\det(\mathbf{I} + \mathbf{u}_1 \mathbf{v}_1^T + \mathbf{u}_2 \mathbf{v}_2^T) = (1 + \mathbf{u}_1^T \mathbf{v}_1)(1 + \mathbf{u}_2^T \mathbf{v}_2) - (\mathbf{u}_1^T \mathbf{v}_2)(\mathbf{u}_2^T \mathbf{v}_1)$ by setting $\mathbf{u}_1 = -s_{k-\tau+u}$, $\mathbf{v}_1 = \hat{\mathcal{B}}_u^k s_{k-\tau+u} / \langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, s_{k-\tau+u} \rangle$, $\mathbf{u}_2 = (\hat{\mathcal{B}}_u^k)^{-1} y_{k-\tau+u}$, and $\mathbf{v}_2 = y_{k-\tau+u} / \langle s_{k-\tau+u}, y_{k-\tau+u} \rangle$. The first inequality follows from (A.3) in Lemma B.2 and the fact $\langle \hat{\mathcal{B}}_u^k s_{k-\tau+u}, s_{k-\tau+u} \rangle \leq \lambda_{\max}(\hat{\mathcal{B}}_u^k) \|s_{k-\tau+u}\|^2$. Actually, we use the fact the trace of a positive definite matrix bounds its maximal eigenvalue for the second inequality. The last inequality follows (A.13). Then, applying (A.12), (A.14) turns to be

$$\det(\hat{\mathcal{B}}_{u+1}^k) \geq \det(\hat{\mathcal{B}}_u^k) \frac{\lambda}{(M + \tau)\mu_2}. \quad (\text{A.15})$$

Applying (A.15) recursively from $u = 0$ to $u = \tau - 1$, we obtain that

$$\det(\hat{\mathcal{B}}_\tau^k) \geq \left[\frac{\lambda}{(M + \tau)\mu_2} \right]^\tau \det(\hat{\mathcal{B}}_0^k).$$

To bound the determinant of $\hat{\mathcal{B}}_0^k$, considering $\hat{\mathcal{B}}_0^k = \text{id}/\gamma_k$ as above and Lemma B.3, we can rewrite for $k \geq 1$ as

$$\det(\hat{\mathcal{B}}_0^k) = \det \left(\frac{\mathbf{I}}{\gamma_k} \right) = \frac{1}{\gamma_k^M} = \left(\frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle} \right)^M \geq \mu_1^M.$$

Consequently, we obtain as

$$\det(\hat{\mathcal{B}}_\tau^k) \geq \mu_1^M \left[\frac{\lambda}{(M+\tau)\mu_2} \right]^\tau.$$

Thus, this yields (A.10), and these complete the proof. \square

Now we prove the main lemma for Proposition 3.1.

Lemma B.7. *Consider the operator $\tilde{\mathcal{H}}^k$ defined by the recursion in (5). Define the constant $0 < \gamma < \Gamma < \infty$. If Assumption 1 holds, the eigenvalues of $\tilde{\mathcal{H}}^k$ is bounded by γ and Γ for all $k \geq 1$ as*

$$\gamma \text{id} \preceq \tilde{\mathcal{H}}^k \preceq \Gamma \text{id}.$$

Proof. The proof is obtained as parallel to the Euclidean case [24]. The sum and product of its eigenvalues of $\hat{\mathcal{B}}^k$ correspond to the bounds on the trace and determinant. Here, we denote λ_i as the i -th largest eigenvalue of the operator matrix $\hat{\mathcal{B}}^k$. From (A.9), the sum of the eigenvalues of $\hat{\mathcal{B}}^k$ satisfies below

$$\sum_{i=1}^M \lambda_i = \text{trace}(\hat{\mathcal{B}}^k) \leq (M+\tau)\mu_2. \quad (\text{A.16})$$

Because all the eigenvalues are positive due to the positive definiteness of $\hat{\mathcal{B}}^k$, it is obvious that every eigenvalues is less than the upper bound of the sum of all of the eigenvalues. Consequently, we obtain $\lambda_i \leq (M+\tau)\mu_2$ for all i , and finally obtain $\hat{\mathcal{B}}^k \preceq (M+\tau)\mu_2 \text{id}$.

On the other hand, because the determinant of a matrix is the product of its eigenvalues, the lower bound in (A.10) bounds the product of the eigenvalues of $\hat{\mathcal{B}}^k$ from below. This means that $\prod_{i=1}^M \lambda_i \geq \frac{\lambda^\tau \mu_1^M}{[(M+\tau)\mu_2]^\tau}$. Thus, we have below for any given eigenvalue of $\hat{\mathcal{B}}^k$, say λ_j ,

$$\lambda_j \geq \frac{1}{\prod_{k=1, k \neq j}^M \lambda_k} \cdot \frac{\lambda^\tau \mu_1^M}{[(M+\tau)\mu_2]^\tau}. \quad (\text{A.17})$$

Considering that $(M+\tau)\mu_2$ is an upper bound for the eigenvalues of $\hat{\mathcal{B}}^k$, $[(M+\tau)\mu_2]^{M-1}$ gives the upper bound of the product of the $(M-1)$ eigenvalues $\prod_{k=1, k \neq j}^M \lambda_k$.

As a result, we obtain that any eigenvalues of $\hat{\mathcal{B}}^k$ is lower bounded as

$$\lambda_j \geq \frac{1}{[(M+\tau)\mu_2]^{M-1}} \cdot \frac{\lambda^\tau \mu_1^M}{[(M+\tau)\mu_2]^\tau} = \frac{\lambda^\tau \mu_1^M}{[(M+\tau)\mu_2]^{M+\tau-1}}. \quad (\text{A.18})$$

Consequently, we finally obtain $\frac{\lambda^\tau \mu_1^M}{[(M+\tau)\mu_2]^{M+\tau-1}} \text{id} \preceq \hat{\mathcal{B}}^k$.

Now, we obtain the claim. The bounds in (A.16) and (A.18) imply that their inverses are bounds for the eigenvalues of $\tilde{\mathcal{H}}^k = (\hat{\mathcal{B}}^k)^{-1}$ as

$$\frac{1}{(M+\tau)\mu_2} \text{id} \preceq \tilde{\mathcal{H}}^k \preceq \frac{[(M+\tau)\mu_2]^{M+\tau-1}}{\lambda^\tau \mu_1^M} \text{id}. \quad (\text{A.19})$$

Denoting $\frac{1}{(M+\tau)\mu_2}$ as γ , and $\frac{[(M+\tau)\mu_2]^{M+\tau-1}}{\lambda^\tau \mu_1^M}$ as Γ , we obtain the claim. This completes the proof. \square

B.2 Main proof of Proposition 3.1

Finally we prove Proposition 3.1.

Proposition 3.1. *Consider the operator $\tilde{\mathcal{H}}^k := \mathcal{T}_{\tilde{\eta}_t^k} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}_{\tilde{\eta}_t^k})^{-1}$, where $\tilde{\mathcal{H}}^k$ is defined by the recursion in (5). Define the constant $0 < \gamma < \Gamma < \infty$. If Assumption 1 holds, the range of eigenvalues of \mathcal{H}_t^k is bounded by γ and Γ for all $k \geq 1, t \geq 1$, i.e.,*

$$\gamma \text{id} \preceq \mathcal{H}_t^k \preceq \Gamma \text{id}. \quad (\text{A.20})$$

Proof. Considering $\mathcal{H}_t^k := \mathcal{T}_{\tilde{\eta}_t^k} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}_{\tilde{\eta}_t^k})^{-1}$, where $\tilde{\eta}_t^k = R_{\tilde{w}^k}^{-1}(w_t^k)$, since $\mathcal{T}_{\tilde{\eta}_t^k}$ is a linear transformation operator, we can conclude that the eigenvalues of \mathcal{H}_t^k and $\tilde{\mathcal{H}}^k$ are identical. Actually, let hat expressions be representation matrices with some bases of $T_{w_t^k} \mathcal{M}$ and $T_{\tilde{w}^k} \mathcal{M}$, we have the relation below;

$$\begin{aligned} \det(\lambda \text{id} - \hat{\mathcal{H}}_t^k) &= \det(\lambda \text{id} - \hat{\mathcal{T}}_{\tilde{\eta}_t^k} \hat{\mathcal{H}}^k (\hat{\mathcal{T}}_{\tilde{\eta}_t^k})^{-1}) \\ &= \det(\hat{\mathcal{T}}_{S_{\tilde{\eta}_t^k}} (\lambda \text{id} - \hat{\mathcal{H}}^k) (\hat{\mathcal{T}}_{\tilde{\eta}_t^k})^{-1}) \\ &= \det(\hat{\mathcal{T}}_{S_{\tilde{\eta}_t^k}}) \det(\lambda \text{id} - \hat{\mathcal{H}}^k) \det((\hat{\mathcal{T}}_{\tilde{\eta}_t^k})^{-1}) \\ &= \det(\hat{\mathcal{T}}_{S_{\tilde{\eta}_t^k}}) \det(\lambda \text{id} - \hat{\mathcal{H}}^k) \det(\hat{\mathcal{T}}_{\tilde{\eta}_t^k})^{-1} \\ &= \det(\lambda \text{id} - \hat{\mathcal{H}}^k). \end{aligned}$$

Therefore, Lemma B.7 directly yields the claim. This completes the proof. \square

C Proof of Theorem 3.2

This section shows the global convergence analysis of the proposed R-SQN-VR. Hereinafter, we use $\mathbb{E}[\cdot]$ to express expectation with respect to the joint distribution of all random variables. For example, w_k is determined by the realizations of the independent random variables $\{z_1, z_2, \dots, z_{k-1}\}$, the total expectation of $f(w_k)$ for any $k \in \mathbb{N}$ can be taken as $\mathbb{E}[f(w_t^k)] = \mathbb{E}_{z_1} \mathbb{E}_{z_2} \dots \mathbb{E}_{z_{k-1}}[f(w_t^k)]$. We also use $\mathbb{E}_{z_t}[\cdot]$ to denote an expected value taken with respect to the distribution of the random variable z_t . This analysis partially extends the expectation based analysis of SGD in the Euclidean [35] into the proposed algorithm.

C.1 Essential lemmas

We first obtain the following lemma from (A.2) in Lemma B.1. Subsequently, $\mathbb{E}_{z_t}[f(w_{t+1}^k)]$ is a meaningful quantity because w_{t+1}^k depends on z_t through the update in Algorithm 1.

Lemma C.1. *Under Lemma B.1, the iterates of Algorithm 1 satisfy the following inequality for all $k \in \mathbb{N}$:*

$$\mathbb{E}_{z_t}[f(w_{t+1}^k)] - f(w_t^k) \leq -\alpha_t^k \langle \text{grad} f(w_t^k), \mathbb{E}_{z_t}[\mathcal{H}_t^k \xi_t^k] \rangle_{w_t^k} + \frac{1}{2} (\alpha_t^k)^2 \Lambda \mathbb{E}_{z_t}[\|\mathcal{H}_t^k \xi_t^k\|_{w_t^k}^2]. \quad (\text{A.21})$$

Proof. When $w_{t+1}^k = R_{w_t^k}(-\alpha_t^k \mathcal{H}_t^k \xi_t^k)$, substituting $-\mathcal{H}_t^k \xi_t^k$ into η_k , the iterates generated by Algorithm 1 satisfy from (A.2) in Lemma B.1

$$\begin{aligned} f(w_{t+1}^k) - f(w_t^k) &\leq \langle \text{grad} f(w_t^k), -\alpha_t^k \mathcal{H}_t^k \xi_t^k \rangle_{w_t^k} + \frac{1}{2} \Lambda \| -\alpha_t^k \mathcal{H}_t^k \xi_t^k \|_{w_t^k}^2 \\ &= -\alpha_t^k \langle \text{grad} f(w_t^k), \mathcal{H}_t^k \xi_t^k \rangle_{w_t^k} + \frac{1}{2} (\alpha_t^k)^2 \Lambda \| \mathcal{H}_t^k \xi_t^k \|_{w_t^k}^2. \end{aligned} \quad (\text{A.22})$$

Taking expectations in the inequalities above with respect to the distribution of z_t , and noting that w_{t+1}^k , but not w_t^k , depends on z_t , we obtain the desired bound. \square

This lemma shows that, regardless of how Algorithm 1 arrived at w_t^k , the expected decrease in the objective function yielded by the k -th step is bounded above by a quantity involving: (i) the *expected directional derivative* of f at w_t^k along $-\mathcal{H}_t^k \xi_t^k$ and (ii) the *second moment* of $\mathcal{H}_t^k \xi_t^k$.

Next, we derive the following lemma;

Lemma C.2. *Under Assumptions 1 and Assumption 2.2, the sequence of average function $f(w_t^k)$ satisfies*

$$\mathbb{E}[f(w_{t+1}^k)] \leq f(w_t^k) - \alpha_t^k \gamma \| \text{grad} f(w_t^k) \|_{w_t^k}^2 + \frac{9\Lambda(\alpha_t^k)^2 \Gamma^2 S^2}{2}. \quad (\text{A.23})$$

Proof. Taking expectation (A.21) in Lemma C.1 with regard to w_t^k considering that \mathcal{H}_t^k is deterministic when w_t^k is given, we write

$$\begin{aligned} &\mathbb{E}[f(w_{t+1}^k)] \\ &\leq f(w_t^k) - \alpha_t^k \langle \text{grad} f(w_t^k), \mathcal{H}_t^k \mathbb{E}_{z_t}[\xi_t^k] \rangle_{w_t^k} + \frac{(\alpha_t^k)^2 \Lambda}{2} \mathbb{E}_{z_t}[\| \mathcal{H}_t^k \xi_t^k \|_{w_t^k}^2]. \\ &\leq f(w_t^k) - \alpha_t^k \langle \text{grad} f(w_t^k), \mathcal{H}_t^k \text{grad} f(w_t^k) \rangle_{w_t^k} + \frac{(\alpha_t^k)^2 \Lambda}{2} \mathbb{E}_{z_t}[\| \mathcal{H}_t^k \xi_t^k \|_{w_t^k}^2]. \\ &\leq f(w_t^k) - \alpha_t^k \langle \text{grad} f(w_t^k), \mathcal{H}_t^k \text{grad} f(w_t^k) \rangle_{w_t^k} + \frac{(\alpha_t^k)^2 \Lambda}{2} \mathbb{E}_{z_t}[\Gamma^2 \| \xi_t^k \|_{w_t^k}^2]. \\ &\leq f(w_t^k) - \alpha_t^k \gamma \| \text{grad} f(w_t^k) \|_{w_t^k}^2 + \frac{9\Lambda(\alpha_t^k)^2 \Gamma^2 S^2}{2}, \end{aligned}$$

where the second inequality is obtained from $\mathbb{E}_{z_t}[\xi_t^k] = \text{grad} f(w_t^k)$ because ξ_t^k is an unbiased estimate of $\text{grad} f(w_t^k)$. The last inequality comes from (11) in Assumption 2.2 since

$$\begin{aligned} \| \xi_t^k \|_{w_t^k} &= \| \text{grad} f_{i_t^k}(w_t^k) - \mathcal{T}_{\tilde{\eta}_t^k}(\text{grad} f_{i_t^k}(\tilde{w}^k)) + \mathcal{T}_{\tilde{\eta}_t^k}(\text{grad} f(\tilde{w}^k)) \|_{w_t^k} \\ &\leq S + S + S = 3S, \end{aligned} \quad (\text{A.24})$$

where $\tilde{\eta}_t^k \in T_{\tilde{w}^k} \mathcal{M}$ satisfies $R_{\tilde{w}^k}(\tilde{\eta}_t^k) = w_t^k$.

This completes the proof. \square

Proposition C.3. *Under Assumptions 1, Assumptions 2, and Lemma C.2, suppose that Algorithm 1 is run with a step-size sequence satisfying (12) in Assumption 2.3. Then, with*

$$A_K := \sum_{k=1}^K \sum_{t=1}^{m_k} \alpha_t^k,$$

$$\mathbb{E} \left[\sum_{k=1}^K \sum_{t=1}^{m_k} \alpha_t^k \|\text{grad} f(w_t^k)\|_{w_t^k}^2 \right] < \infty \quad (\text{A.25})$$

$$\text{and therefore } \mathbb{E} \left[\frac{1}{A_K} \sum_{k=1}^K \sum_{t=1}^{m_k} \alpha_t^k \|\text{grad} f(w_t^k)\|_{w_t^k}^2 \right] \xrightarrow{K \rightarrow \infty} 0 \quad (\text{A.26})$$

Proof. Taking the total expectation of (A.23) in Lemma C.2 yields

$$\mathbb{E}[f(w_{t+1}^k)] - \mathbb{E}[f(w_t^k)] \leq -\alpha_t^k \gamma \mathbb{E}[\|\text{grad} f(w_t^k)\|_{w_t^k}^2] + \frac{9\Lambda(\alpha_t^k)^2 \Gamma^2 S^2}{2}.$$

Summing both sides of this inequality for $\{w_1^1, \dots, w_{m_1}^1, \dots, w_1^{K-1}, \dots, w_{m_{K-1}}^{K-1}, w_1^K, \dots, w_{m_K}^K\}$ gives

$$\begin{aligned} f_{\inf} - f(w_1^1) &\leq \mathbb{E}[f(w_{k+1}^k)] - f(w_1^1) \\ &\leq -\gamma \sum_{k=1}^K \sum_{t=1}^{m_k} \alpha_t^k \mathbb{E}[\|\text{grad} f(w_t^k)\|_{w_t^k}^2] + \frac{9\Lambda \Gamma^2 S^2}{2} \sum_{k=1}^K \sum_{t=1}^{m_k} (\alpha_t^k)^2. \end{aligned}$$

Dividing by γ and rearranging the terms, we obtain

$$\sum_{k=1}^K \sum_{t=1}^{m_k} \alpha_t^k \mathbb{E}[\|\text{grad} f(w_t^k)\|_{w_t^k}^2] \leq \frac{(f(w_1^1) - f_{\inf})}{\gamma} + \frac{9\Lambda \Gamma^2 S^2}{2\gamma} \sum_{k=1}^K \sum_{t=1}^{m_k} (\alpha_t^k)^2.$$

The second condition in (12) in Assumption 2.3 implies that the right-hand side of this inequality converges to a *finite limit* when K increases, proving (A.25). Then, (A.26) follows since the first condition in (12) of Assumption 2.3 ensures that $A_K \rightarrow \infty$ as $K \rightarrow \infty$. \square

Proposition C.3 states about a *weighted sum-of-squares* and a *weighted average of squared gradients* of f . In particular, (A.26) concludes that the weighted average norm of the squared gradients converges to zero even if the gradient are noisy. But, the fact only specifies a property of a weighted average is only of minor importance since one can still conclude *the expected gradient norms cannot asymptotically stay far from zero*.

Then, we obtain the following proposition by taking (A.25) into account with the first condition of (12) of Assumption 2.3.

Proposition C.4. *Under Assumptions 1, 2, and Lemma C.2, suppose that Algorithm 1 is run with a step-size sequence satisfying (12) in Assumption 2.3. Then,*

$$\liminf_{k \rightarrow \infty} \mathbb{E}[\|\text{grad} f(w_t^k)\|_{w_t^k}^2] = 0. \quad (\text{A.27})$$

Proof. By contradiction. Assume that (A.27) does not hold. Then, there exists $\delta > 0$ such that $\mathbb{E}[\|\text{grad} f(w_t^k)\|_{w_t^k}^2] > \delta$ for all k sufficiently large, say, $k > N$. We have

$$\mathbb{E} \left[\sum_{k=N}^{\infty} \sum_{t=1}^{m_k} \alpha_t^k \|\text{grad} f(w_t^k)\|_{w_t^k}^2 \right] \geq \sum_{k=N}^{\infty} \sum_{t=1}^{m_k} \alpha_t^k \mathbb{E}[\|\text{grad} f(w_t^k)\|_{w_t^k}^2] > \delta \sum_{k=N}^{\infty} \sum_{t=1}^{m_k} \alpha_t^k = \infty.$$

This contradicts (A.25). \square

A “lim inf” result of this type should be familiar to those knowledgeable of the nonlinear optimization literature. This intuition is that, for the R-SGD with diminishing step-sizes, the expected gradient norms cannot stay bounded away from zero.

C.2 Main proof of Theorem 3.2

Theorem. 3.2. *Consider Algorithm 1 and suppose Assumptions 1 and 2, and that the mapping $w \mapsto \|\text{grad}f(w)\|_w^2$ has the positive real number that the largest eigenvalue of its Riemannian Hessian is bounded by for all $w \in \mathcal{M}$. Then, we have*

$$\lim_{k \rightarrow \infty} \mathbb{E}[\|\text{grad}f(w_t^k)\|_{w_t^k}^2] = 0.$$

Proof. We define $h(w)$ as $h(w) := \|\text{grad}f(w)\|_{w_t^k}^2$ and let Λ_h be the absolute value of the eigenvalue with the largest magnitude of the Hessian of h . Then, from Taylor’s theorem, we obtain

$$h(w_{t+1}^k) - h(w_t^k) \leq -2\alpha_t^k \langle \text{grad}h(w_t^k), \text{Hess}f(w_t^k)[\mathcal{H}_t^k \xi_t^k] \rangle_{w_t^k} + \frac{1}{2}(\alpha_t^k)^2 \Lambda_h \|\mathcal{H}_t^k \xi_t^k\|_{w_t^k}^2.$$

Taking the expectation with respect to the distribution of z_t , we obtain below;

$$\begin{aligned} & \mathbb{E}_{z_t}[h(w_{t+1}^k)] - h(w_t^k) \\ & \leq -2\alpha_t^k \langle \text{grad}f(w_t^k), \mathbb{E}_{z_t}[\text{Hess}f(w_t^k)[\mathcal{H}_t^k \xi_t^k]] \rangle_{w_t^k} + \frac{1}{2}(\alpha_t^k)^2 \Lambda_h \mathbb{E}_{z_t}[\|\mathcal{H}_t^k \xi_t^k\|_{w_t^k}^2] \\ & = -2\alpha_t^k \langle \text{grad}f(w_t^k), \text{Hess}f(w_t^k)[\mathcal{H}_t^k \mathbb{E}_{z_t}[\xi_t^k]] \rangle_{w_t^k} + \frac{1}{2}(\alpha_t^k)^2 \Lambda_h \mathbb{E}_{z_t}[\|\mathcal{H}_t^k \xi_t^k\|_{w_t^k}^2] \\ & \leq 2\alpha_t^k \|\text{grad}f(w_t^k)\|_{w_t^k} \|\text{Hess}f(w_t^k)[\mathcal{H}_t^k \mathbb{E}_{z_t}[\xi_t^k]]\|_{w_t^k} + \frac{1}{2}(\alpha_t^k)^2 \Lambda_h \Gamma^2 \mathbb{E}_{z_t}[\|\xi_t^k\|_{w_t^k}^2] \\ & \leq 2\alpha_t^k \Lambda \Gamma \|\text{grad}f(w_t^k)\|_{w_t^k}^2 + \frac{9}{2}(\alpha_t^k)^2 \Lambda_h S^2 \Gamma^2, \end{aligned}$$

where the last inequality comes from $\|\text{Hess}f(w)[\mathcal{H}_t^k \text{grad}f(w_t^k)]\|_{w_t^k} \leq \Lambda \|\mathcal{H}_t^k \text{grad}f(w_t^k)\|_{w_t^k} \leq \Lambda \Gamma \|\text{grad}f(w_t^k)\|_{w_t^k}$.

Taking the total expectation simply yields

$$\mathbb{E}[h(w_{t+1}^k)] - \mathbb{E}[h(w_t^k)] \leq 2\alpha_t^k \Lambda \Gamma \mathbb{E}[\|\text{grad}f(w_t^k)\|_{w_t^k}^2] + \frac{9}{2}(\alpha_t^k)^2 \Lambda_h S^2 \Gamma^2. \quad (\text{A.28})$$

Recall that Proposition C.3 establishes that the first component of this bound is the term of a convergent sum. The second component of this bound is also the term of a convergent sum since $\sum_{k=1}^{\infty} \sum_{t=1}^{m_k} (\alpha_t^k)^2$ converges. This means that again the result of Proposition C.3 can be applied. Therefore, the right-hand side of (A.28) is the term of a convergent sum. Let us now define $S_K^+ = \sum_{k=1}^K \sum_{t=1}^{m_k} \max(0, \mathbb{E}[h(w_{t+1}^k)] - \mathbb{E}[h(w_t^k)])$, and $S_K^- = \sum_{k=1}^K \sum_{t=1}^{m_k} \max(0, \mathbb{E}[h(w_t^k)] - \mathbb{E}[h(w_{t+1}^k)])$.

Since the bound (A.28) is positive and forms a convergent sum, the nondecreasing sequence S_K^+ is upper bounded and therefore converges. Since, for any $K \in \mathbb{N}$, one has $\mathbb{E}[h(w_K)] = h(w_1) + S_K^+ - S_K^- \geq 0$, the nondecreasing sequence S_K^- is upper bounded and therefore also converges. Therefore $\mathbb{E}[h(w_K)]$ converges. Consequently, this implies that this limit must be zero from Proposition C.4. This completes the proof. \square

D Proof of Theorem 3.3

This section first introduces some essential lemmas. Then, the main proof of Theorem 3.3 is given. This section also derives at the end a corollary about the analysis when the using exponential mapping and the parallel translation that are special cases of the retraction and the vector transport.

D.1 Essential lemmas

We first introduce the following lemmas.

Lemma D.1 (In the proof of Lemma 3.9 in [20]). *Under Assumptions 1.1 and 1.2, there exists a constant $\beta > 0$ such that*

$$\|P_\gamma^{w \leftarrow z}(\text{grad}f(z)) - \text{grad}f(w)\|_w \leq \beta \text{dist}(z, w), \quad (\text{A.29})$$

where w and z are in Θ in Assumption 1.2 and γ is a curve $\gamma(t) := R_z(\tau\eta)$ for $\eta \in T_z\mathcal{M}$ defined by a retraction R on \mathcal{M} . $P_\gamma^{w \leftarrow z}(\cdot)$ is a parallel translation operator along the curve γ from z to w .

Note that the curve γ in this lemma is not necessarily the geodesic. The relation (A.29) is a generalization of the Lipschitz continuity condition.

Lemma D.2 (Lemma 3.5 in [20]). *Let $\mathcal{T} \in C^0$ be a vector transport associated with the same retraction R as that of the parallel translation $P \in C^\infty$. Under Assumption 1.5, for any $\bar{w} \in \mathcal{M}$ there exists a constant $\theta > 0$ and a neighborhood \mathcal{U} of \bar{w} such that for all $w, z \in \mathcal{U}$,*

$$\|\mathcal{T}_\eta \xi - P_\eta \xi\|_z \leq \theta \|\xi\|_w \|\eta\|_w, \quad (\text{A.30})$$

where $\xi, \eta \in T_w\mathcal{M}$ and $R_w(\eta) = z$.

Modifying slightly Lemma 3 in [36], we have the following lemma.

Lemma D.3 (Lemma 3 in [36]). *Let \mathcal{M} be a Riemannian manifold endowed with retraction R and let $\bar{w} \in \mathcal{M}$. Then there exist $\tau_1 > 0$, $\tau_2 > 0$ and δ_{τ_1, τ_2} such that for all w in a sufficiently small neighborhood of \bar{w} and all $\xi \in T_w\mathcal{M}$ with $\|\xi\|_w \leq \delta_{\tau_1, \tau_2}$, the inequalities*

$$\tau_1 \text{dist}(w, R_w(\xi)) \leq \|\xi\|_w \leq \tau_2 \text{dist}(w, R_w(\xi)) \quad (\text{A.31})$$

hold.

We also show a property of the Karcher mean on a general Riemannian manifold.

Lemma D.4 (Lemma C.2 in [21]). *Let w_1, \dots, w_m be points on a Riemannian manifold \mathcal{M} and let w be the Karcher mean of the m points. For an arbitrary point p on \mathcal{M} , we have*

$$(\text{dist}(p, w))^2 \leq \frac{4}{m} \sum_{i=1}^m (\text{dist}(p, w_i))^2.$$

We now bound the variance of ξ_t^k as follows.

Lemma D.5 (Lemma 5.8 in [21]). *Suppose Assumptions 1.1, 1.2, 1.4, and 1.5, which guarantees Lemmas D.1, D.2, and D.3 for $\bar{w} = w^*$. Let $\beta > 0$ be a constant such that*

$$\|P_\gamma^{w \leftarrow z}(\text{grad} f_n(z)) - \text{grad} f_n(w)\|_w \leq \beta \text{dist}(z, w), \quad w, z \in \Theta, \quad n = 1, 2, \dots, N.$$

The existence of such β is guaranteed by Lemma D.1. The upper bound of the variance of ξ_t^k is given by

$$\mathbb{E}_{i_t^k}[\|\xi_t^k\|_{w_t^k}^2] \leq 4(\beta^2 + \tau_2^2 C^2 \theta^2)(7(\text{dist}(w_t^k, w^*))^2 + 4(\text{dist}(\tilde{w}^k, w^*))^2), \quad (\text{A.32})$$

where the constant θ corresponds to that in Lemma D.2, C is the upper bound of $\|\text{grad} f_n(w)\|_w$, $n = 1, 2, \dots, N$ for $w \in \Theta$, and $\tau_2 > 0$ appears in (A.31).

We also have the following corollary of the previous lemma with the case $R = \text{Exp}$ and $\mathcal{T} = P$.

Corollary D.6 (Corollary 5.1 in [21]). *Consider Algorithm 1 with $\mathcal{T} = P$ and $R = \text{Exp}$, i.e., the parallel translation and the exponential mapping case. When each $\text{grad} f_n$ is β_0 -Lipschitz continuously differentiable, the upper bound of the variance of ξ_t^k is given by*

$$\mathbb{E}_{i_t^k}[\|\xi_t^k\|_{w_t^k}^2] \leq \beta_0^2(14(\text{dist}(w_t^k, w^*))^2 + 8\text{dist}(\tilde{w}^k, w^*))^2). \quad (\text{A.33})$$

Next, we show the lemma that finds a lower bound for $\|\text{grad} f(w_t^k)\|_{w_t^k}$ with respect to the error $f(w_t^k) - f(w^*)$, which is a standard derivation in the Euclidean space. See, e.g., [37]. We extend this into manifolds.

Lemma D.7. *Let $w \in \mathcal{M}$ and z be in a totally retractive neighborhood of w . It holds that*

$$2\lambda(f(w) - f(z)) \leq \|\text{grad} f(w)\|_w^2. \quad (\text{A.34})$$

Proof. Let $\zeta = R_w^{-1}(z)$. Using (A.1) in Lemma B.1, which is equivalent to the strong convexity of $\text{grad} f$, we obtain

$$\begin{aligned} f(z) &\geq f(w) + \langle \text{grad} f(w), \zeta \rangle_w + \frac{\lambda}{2} \|\zeta\|_w^2 \\ &\geq f(w) + \min_{\xi \in T_w \mathcal{M}} \left(\langle \text{grad} f(w), \xi \rangle_w + \frac{\lambda}{2} \|\xi\|_w^2 \right) \\ &\geq f(w) - \frac{1}{2\lambda} \|\text{grad} f(w)\|_w^2. \end{aligned}$$

Rearranging this inequality completes the proof. \square

D.2 Main proof of Theorem 3.3

Theorem 3.3. *Let \mathcal{M} be a Riemannian manifold and $w^* \in \mathcal{M}$ be a non-degenerate local minimizer of f (i.e., $\text{grad} f(w^*) = 0$ and the Hessian $\text{Hess} f(w^*)$ of f at w^* is positive definite). Suppose Assumption 1 holds. Let the constants β, θ , and C be in Lemma D.5, and τ_1 and τ_2 be in Lemma D.3. λ and Λ are the constants in Lemma B.1. γ and Γ are constants in Proposition 3.1. Let α be a positive number satisfying $\lambda\tau_1^2 > 2\alpha(\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2 C^2 \theta^2))$ and $\gamma\lambda^2\tau_1^2 > 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2 C^2 \theta^2)$. It then follows that for any sequence $\{\tilde{w}^k\}$ generated by*

Algorithm 1 under a fixed step-size $\alpha_t^k := \alpha$ and $m_k := m$ converging to w^* , there exists $K > 0$ such that for all $k > K$,

$$\mathbb{E}[(\text{dist}(\tilde{w}^{k+1}, w^*))^2] \leq \frac{2(\Lambda\tau_2^2 + 16m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))}{m\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))} \mathbb{E}[(\text{dist}(\tilde{w}^{k-1}, w^*))^2].$$

Proof. Using (A.2) in Lemma B.1, which is equivalent to the Lipschitz continuity of $\text{grad}f$ from Assumption 3, we obtain

$$f(w_{t+1}^k) - f(w_t^k) \leq \langle \text{grad}f(w_t^k), -\alpha\mathcal{H}_t^k\xi_t^k \rangle_{w_t^k} + \frac{1}{2}\Lambda(-\alpha\|\mathcal{H}_t^k\xi_t^k\|_{w_t^k})^2.$$

Taking expectation with regard to z_t , this becomes

$$\begin{aligned} \mathbb{E}_{z_t}[f(w_{t+1}^k)] - f(w_t^k) &\leq \mathbb{E}_{z_t}[\langle \text{grad}f(w_t^k), -\alpha\mathcal{H}_t^k\xi_t^k \rangle_{w_t^k} + \frac{1}{2}\alpha^2\Lambda\|\mathcal{H}_t^k\xi_t^k\|_{w_t^k}^2] \\ &\leq -\alpha\langle \text{grad}f(w_t^k), \mathbb{E}_{z_t}[\mathcal{H}_t^k\xi_t^k] \rangle_{w_t^k} + \frac{1}{2}\alpha^2\Lambda\mathbb{E}_{z_t}[\|\mathcal{H}_t^k\xi_t^k\|_{w_t^k}^2] \\ &\leq -\alpha\langle \text{grad}f(w_t^k), \mathcal{H}_t^k\text{grad}f(w_t^k) \rangle_{w_t^k} + \frac{1}{2}\alpha^2\Lambda\mathbb{E}_{z_t}[\|\mathcal{H}_t^k\xi_t^k\|_{w_t^k}^2] \\ &\leq -\alpha\gamma\|\text{grad}f(w_t^k)\|_{w_t^k}^2 + \frac{1}{2}\alpha^2\Lambda\Gamma^2\mathbb{E}_{z_t}[\|\xi_t^k\|_{w_t^k}^2]. \end{aligned} \quad (\text{A.35})$$

where the third inequality used the fact that $\mathbb{E}_{z_t}[\mathcal{H}_t^k\xi_t^k] = \mathcal{H}_t^k\text{grad}f(w_t^k)$. The last inequality used the bound of \mathcal{H}_t^k in Proposition 3.1.

From Lemma D.7, (A.35) yields

$$\mathbb{E}_{z_t}[f(w_{t+1}^k)] - f(w_t^k) \leq -2\alpha\gamma\lambda(f(w_t^k) - f(w^*)) + \frac{1}{2}\alpha^2\Lambda\Gamma^2\mathbb{E}_{z_t}[\|\xi_t^k\|_{w_t^k}^2]. \quad (\text{A.36})$$

Using (A.1) in Lemma B.1 with $\text{grad}f(w^*) = 0$, and using Lemma D.3, we obtain

$$f(w_t^k) - f(w^*) \geq \frac{\lambda}{2}\|R_{w^*}^{-1}(w_t^k)\|_{w^*}^2 \geq \frac{\lambda\tau_1^2}{2}(\text{dist}(w_t^k, w^*))^2. \quad (\text{A.37})$$

Plugging (A.37) and the bound of $\mathbb{E}_{i_t^k}[\|\xi_t^k\|^2]$ in (A.32) in Lemma D.5 into (A.36) yields

$$\begin{aligned} \mathbb{E}_{z_t}[f(w_{t+1}^k)] - f(w_t^k) &\leq -\alpha\gamma\lambda^2\tau_1^2(\text{dist}(w_t^k, w^*))^2 + \frac{1}{2}\alpha^2\Lambda\Gamma^2\mathbb{E}_{z_t}[\|\xi_t^k\|_{w_t^k}^2] \\ &\leq -\alpha\gamma\lambda^2\tau_1^2(\text{dist}(w_t^k, w^*))^2 \\ &\quad + \frac{1}{2}\alpha^2\Lambda\Gamma^2\{4(\beta^2 + \tau_2^2C^2\theta^2)(7(\text{dist}(w_t^k, w^*))^2 + 4(\text{dist}(\tilde{w}^k, w^*))^2)\} \\ &\leq (-\alpha\gamma\lambda^2\tau_1^2 + 14\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))(\text{dist}(w_t^k, w^*))^2 \\ &\quad + 8\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)(\text{dist}(\tilde{w}^k, w^*))^2. \end{aligned}$$

Taking expectations over all random variables, we obtain below by further summing over $t = 0, \dots, m-1$ of the inner loop on k -th epoch

$$\begin{aligned} \mathbb{E}[f(w_m^k) - f(w_0^k)] &\leq -(\alpha\gamma\lambda^2\tau_1^2 - 14\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)) \sum_{t=0}^{m-1} \mathbb{E}[(\text{dist}(w_t^k, w^*))^2] \\ &\quad + 8m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)\mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2]. \end{aligned} \quad (\text{A.38})$$

Here, considering the difference with the solution w^* in terms of the cost function value, we obtain

$$\begin{aligned}\mathbb{E}[f(w_m^k) - f(w_0^k)] &= \mathbb{E}[f(w_m^k) - f(w^*) - (f(w_0^k) - f(w^*))] \\ &\geq \frac{1}{2}\mathbb{E}[\lambda\tau_1^2(\text{dist}(w_m^k, w^*))^2 - \Lambda\tau_2^2(\text{dist}(w_0^k, w^*))^2].\end{aligned}$$

Plugging the above into (A.38) yields

$$\begin{aligned}&\mathbb{E}[\lambda\tau_1^2(\text{dist}(w_m^k, w^*))^2 - \Lambda\tau_2^2(\text{dist}(w_0^k, w^*))^2] \\ &\leq -2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)) \sum_{t=0}^{m-1} \mathbb{E}[(\text{dist}(w_t^k, w^*))^2] \\ &\quad + 16m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)\mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2].\end{aligned}$$

Rearranging this gives

$$\begin{aligned}&2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)) \sum_{t=0}^{m-1} \mathbb{E}[(\text{dist}(w_t^k, w^*))^2] \\ &\leq \mathbb{E}[\Lambda\tau_2^2(\text{dist}(w_0^k, w^*))^2] - \mathbb{E}[\lambda\tau_1^2(\text{dist}(w_m^k, w^*))^2] \\ &\quad + 16m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)\mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2].\end{aligned}\tag{A.39}$$

Now, addressing option I in Algorithm 1, which uses $\tilde{w}^{k+1} = g_{m_k}(w_1^k, \dots, w_m^k)$, we derive below from Lemma D.4 as

$$\begin{aligned}&\frac{m}{4}2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))\mathbb{E}[(\text{dist}(\tilde{w}^{k+1}, w^*))^2] \\ &\leq 2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))\mathbb{E}\left[\sum_{t=0}^{m-1} (\text{dist}(w_t^k, w^*))^2 + (\text{dist}(w_m^k, w^*))^2 - (\text{dist}(w_0^k, w^*))^2\right] \\ &\stackrel{(A.39)}{\leq} \mathbb{E}[\Lambda\tau_2^2(\text{dist}(w_0^k, w^*))^2] - \mathbb{E}[\lambda\tau_1^2(\text{dist}(w_m^k, w^*))^2] + 16m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)\mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2] \\ &\quad + 2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))\mathbb{E}[(\text{dist}(w_m^k, w^*))^2 - (\text{dist}(w_0^k, w^*))^2] \\ &\leq (\Lambda\tau_2^2 - 2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)))\mathbb{E}[(\text{dist}(w_0^k, w^*))^2] \\ &\quad + 16m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)\mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2] \\ &\quad - (\lambda\tau_1^2 - 2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)))\mathbb{E}[(\text{dist}(w_m^k, w^*))^2].\end{aligned}$$

Combining the relation $\Lambda\tau_2^2 > \lambda\tau_1^2$ and the assumption $\lambda\tau_1^2 > 2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))$, since $w_0^k = \tilde{w}^k$, we obtain

$$\begin{aligned}&\frac{m}{4}2\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))\mathbb{E}[(\text{dist}(\tilde{w}^{k+1}, w^*))^2] \\ &\leq (\Lambda\tau_2^2 + 16m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))\mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2].\end{aligned}$$

Finally, we obtain

$$\mathbb{E}[(\text{dist}(\tilde{w}^{k+1}, w^*))^2] \leq \frac{2(\Lambda\tau_2^2 + 16m\alpha^2\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))}{m\alpha(\gamma\lambda^2\tau_1^2 - 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2))} \mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2]\tag{A.40}$$

This completes the proof. \square

Remark D.8. From the proof of Lemma B.4, if we adopt the parallel translation as the vector transport, i.e., $\mathcal{T} = P$, the first two terms in (A.6) are equal to zero, and μ_2 in (A.5) gets smaller than that of the case of vector transport. This leads to a smaller Γ and a larger γ in (10) of Proposition 3.1. Then, the smaller Γ and the larger γ in (10) leads to a smaller coefficient in (A.40) of Theorem 3.3. Consequently, the parallel translation can result in a faster local convergence rate.

We obtain the following corollary of the previous theorem with the case $R = \text{Exp}$ and $\mathcal{T} = P$.

Corollary D.9. Consider Algorithm 1 with $\mathcal{T} = P$ and $R = \text{Exp}$, i.e., the parallel translation and the exponential mapping case. Let \mathcal{M} be a Riemannian manifold and $w^* \in \mathcal{M}$ be a non-degenerate local minimizer of f (i.e., $\text{grad}f(w^*) = 0$ and the Hessian $\text{Hess}f(w^*)$ of f at w^* is positive definite). Suppose Assumptions 1 hold. Let the constants θ , and C in Lemma D.5. β_0 is the constant in Corollary D.6. λ and Λ are the constants in Lemma B.1 satisfying (A.1). γ and Γ are constants in Theorem 3.1. Let α be a positive number satisfying $\lambda > 2\alpha(\gamma\lambda^2 - 7\alpha\Lambda\Gamma^2\beta_0^2)$ and $\gamma\lambda^2\tau_1^2 > 14\alpha\Lambda\Gamma^2(\beta^2 + \tau_2^2C^2\theta^2)$. It then follows that for any sequence $\{\tilde{w}^k\}$ generated by Algorithm 1 with a fixed step-size $\alpha_t^k := \alpha$ and $m_k := m$ converging to w^* , there exists $K > 0$ such that for all $k > K$,

$$\mathbb{E}[(\text{dist}(\tilde{w}^{k+1}, w^*))^2] \leq \frac{2(\Lambda + 8m\alpha^2\Lambda\Gamma^2\beta_0^2)}{m\alpha(\gamma\lambda^2 - 7\alpha\Lambda\Gamma^2\beta_0^2)}\mathbb{E}[(\text{dist}(\tilde{w}^k, w^*))^2] \quad (\text{A.41})$$

Proof. The proof is given similarly to Theorem 3.3. We use Corollary D.6, and also set as $\theta = 0$ in Lemma D.2, and as $\tau_1 = \tau_2 = 1$ in Lemma D.3. \square

E Isometric vector transport \mathcal{S}_{++}^d

This section introduces the isometric vector transport, which fulfills the locking condition for \mathcal{S}_{++}^d . [20, 29] describe the details, and the comprehensive explanations therein.

Denoting B as the function giving a basis of $T_w\mathcal{M}$ as $B : w \rightarrow B(w) = (b_1, b_2, \dots, b_d)$, where d denotes the dimension of \mathcal{M} . b_i is i -th orthonormal basis of $T_w\mathcal{M}$. Here, we consider $w \in \mathcal{M}$, $\eta, \xi \in T_w\mathcal{M}$, $z = R_w(\eta)$, $B_1 = B(w)$, and $B_2 = B(z)$. Then, the isometric vector transport is defined as

$$\mathcal{T}_\eta\xi = B_2 \left(\mathbf{I} - \frac{2v_2v_2^T}{v_2^Tv_2} \right) \left(\mathbf{I} - \frac{2v_1v_1^T}{v_1^Tv_1} \right) B_1^b\xi, \quad (\text{A.42})$$

where $v_1 = B_1^b\eta - y$, $v_2 = y - \beta B_2^bT_{R_\eta}\eta$. y can be any vector satisfying $\|y\| = \|B_1^b\eta\| = \|\beta B_2^bT_{R_\eta}\eta\|$. The orthonormal basis of $T_{\mathbf{X}}\mathcal{M}$ is defined by $B_{\mathbf{X}} = \{\mathbf{L}e_ie_i^T\mathbf{L}^T : i = 1, \dots, n\} \cup \{1/\sqrt{2}\mathbf{L}(e_ie_j^T + e_je_i^T)\mathbf{L}^T, i < j, i = 1, \dots, n, j = 1, \dots, n\}$. Here, $\{e_1, \dots, e_n\}$ is the canonical basis of Euclidean space of which dimension is n . $\mathbf{L}\mathbf{L}^T$ represents the Cholesky decomposition of \mathbf{X} .

F Additional numerical experiments

In this section, we show additional numerical experiments which do not appear in the main text.

F.1 Karcher mean problem

We consider the PSD Karcher mean problem of $N = 1000$, $d = 3$ (**Case KM-1: small size instance**). Figure A.1 shows the convergence plots of the cost function, the optimality gap and the norm of the gradient for all the five runs. They show that the proposed R-SQN-VR gives a better performance beyond other stochastic algorithms. Additionally, we consider the case of $N = 10000$, $d = 10$ (**Case KM-2: large size instance**). From Figure A.2, we find that our proposed algorithm R-SQN-VR gives a better performance than other stochastic algorithms.

F.2 Matrix completion problem on synthetic datasets

This section shows the results of six problem instances. Due to the page limitations, we only show the loss on a test set Φ , which is different from the training set Ω . The loss on the test set demonstrates the convergence speed to a good prediction accuracy of missing entries.

Case MC-S1: We first show the results of the comparison when the number of samples $N = 5000$, the dimension $d = 500$, the memory size $L = 5$, the oversampling ratio (OS) is 6, and the condition number (CN) 5. We also add Gaussian noise $\sigma = 10^{-10}$. Figures A.3 show the results of five runs. They show superior performances than other algorithms.

Case MC-S2: influence on low sampling. We look into problem instances from scarcely sampled data, e.g. OS is 4. Other conditions are the same as **Case MC-S1**. From Figures A.4, we can find that the proposed algorithm gives much better and stabler performances against other algorithms.

Case MC-S3: influence on ill-conditioning. We consider the problem instances with higher condition number (CN) 10. Other conditions are the same as **Case MC-S1**. Figures A.5 show the superior performances of the proposed algorithm against other algorithms.

Case MC-S4: influence on large memory size. We evaluate the performance when the memory size L is large, where $L = 10$. Other conditions are the same as **Case MC-S1**. From Figures A.6, the proposed R-SQN-VR still shows the superior performances against other algorithms.

Case MC-S5: influence on higher noise. We consider noisy problem instances, where $\sigma = 10^{-6}$. Other conditions are the same as **Case MC-S1**. Figures A.7 show that the convergent MSE values are much higher than the other cases. Then, we can see the superior performance of the proposed R-SQN-VR against other algorithms.

Case MC-S6: influence on higher rank. We consider problem instances with higher rank, where $r = 10$. Other conditions are the same as **Case MC-S1**. From Figures A.8, the proposed R-SQN-VR still shows the superior performances against other algorithms. Especially, the exponential mapping and the parallel translation case of the proposed method gives much better performance than that of the retraction and the vector transport. In addition, Grouse indicates the faster decrease of the MSE at the begging of the iterations. However, the convergent MSE values are much higher than those of others.

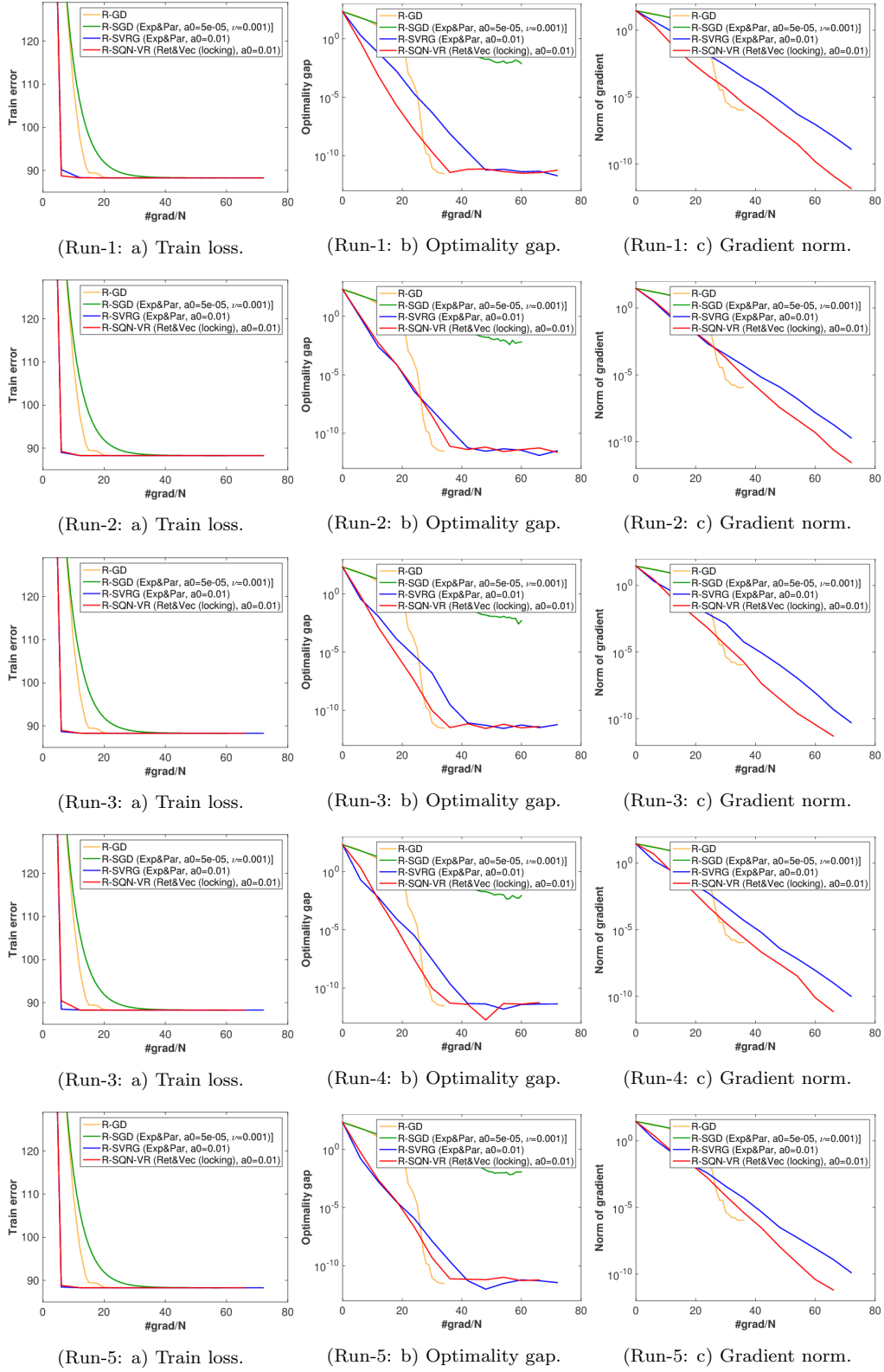


Figure A.1: Performance evaluations on Karcher mean problem (**Case KM-1: small size instance**).

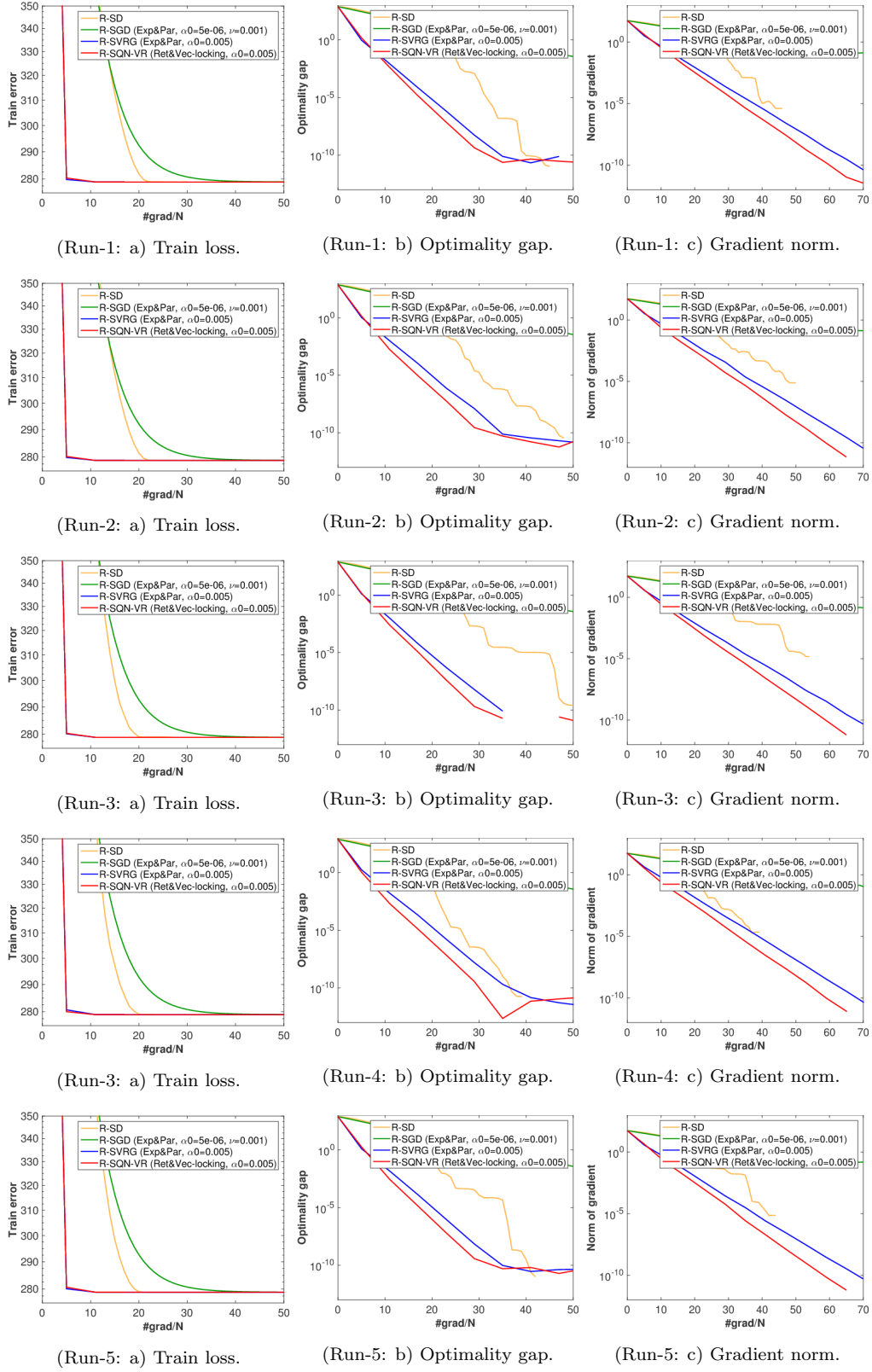


Figure A.2: Performance evaluations on Karcher mean problem (**Case KM-2: large size instance**).

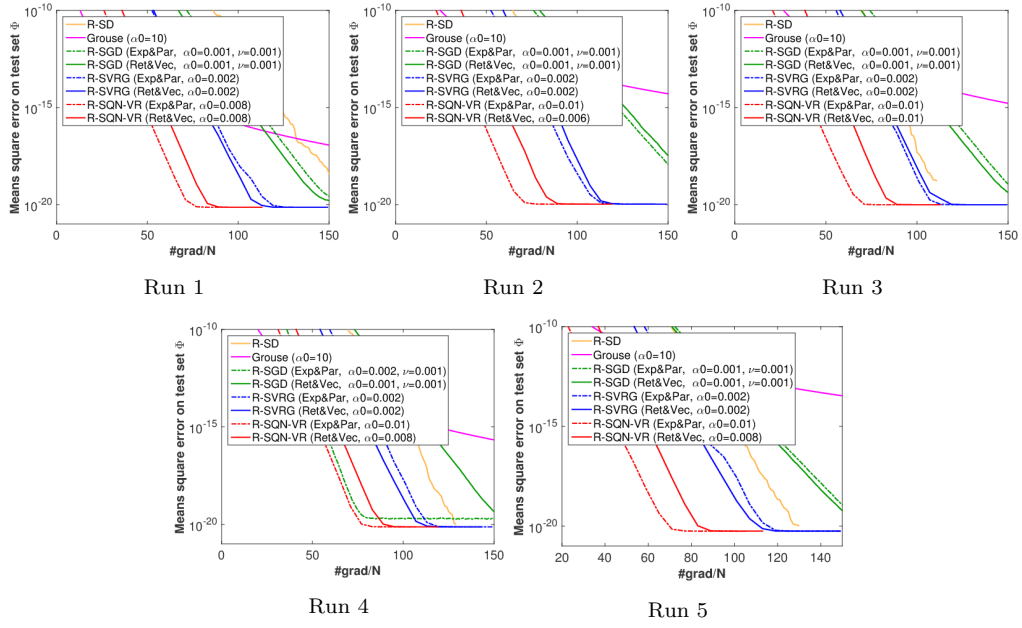


Figure A.3: Performance evaluations on low-rank matrix completion problem (Case MC-S1: baseline).

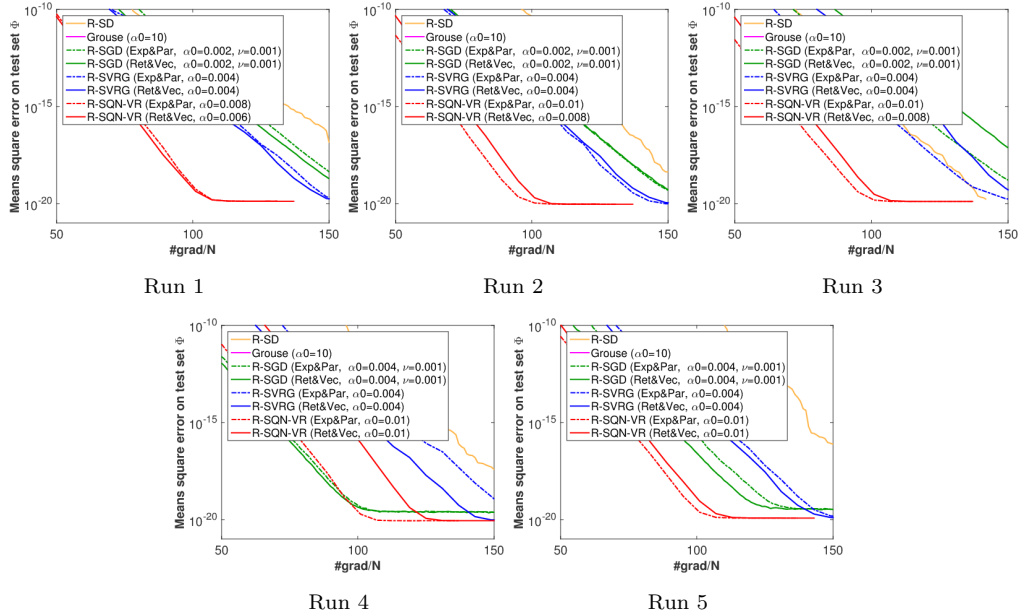


Figure A.4: Performance evaluations on low-rank matrix completion problem (Case MC-S2: influence on lower sampling).

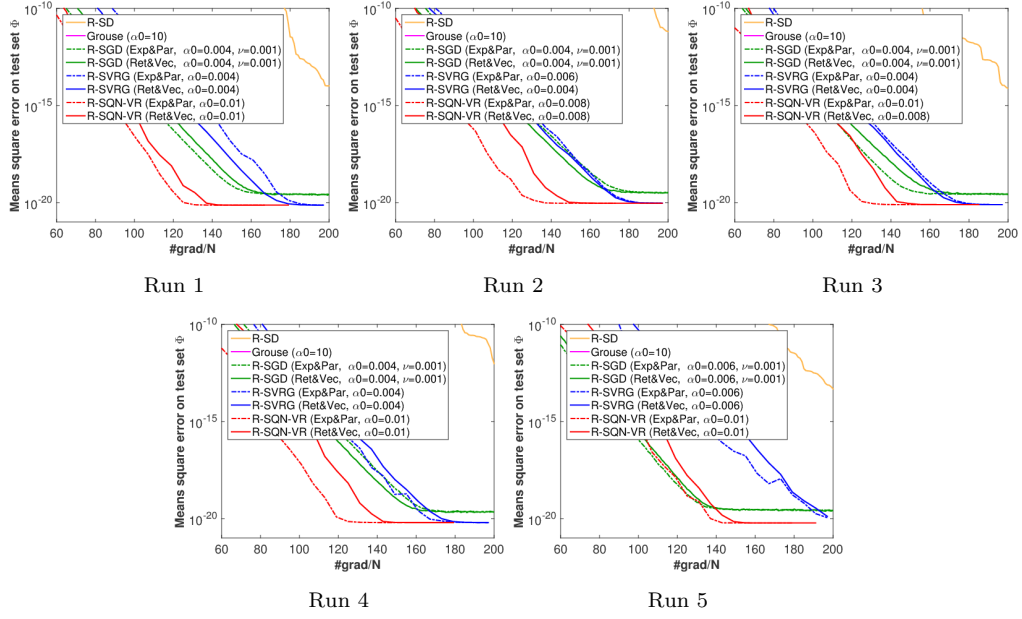


Figure A.5: Performance evaluations on low-rank matrix completion problem (Case MC-S3: influence on ill-conditioning).

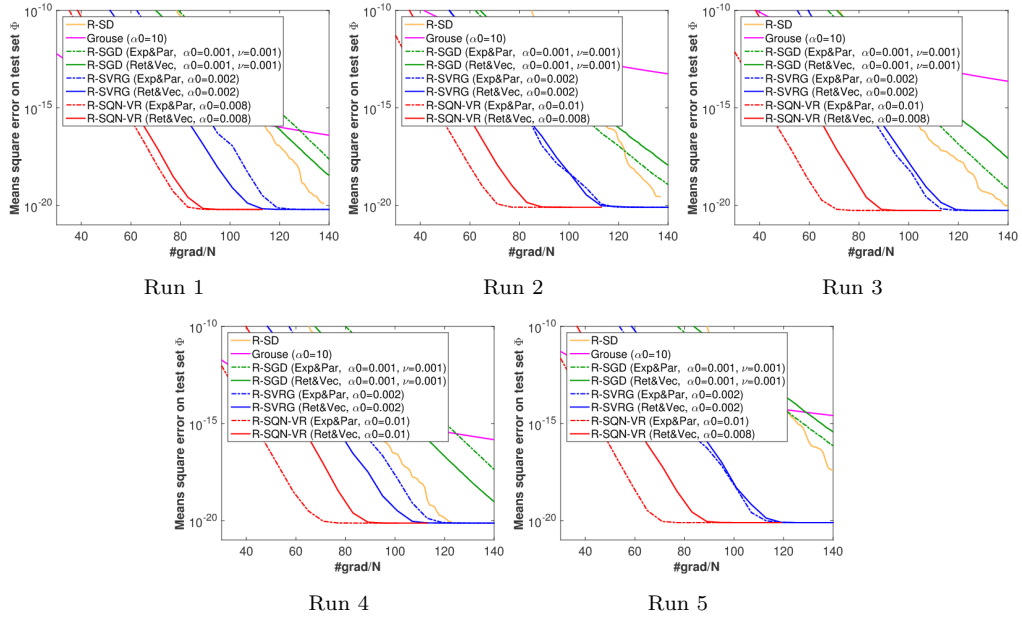


Figure A.6: Performance evaluations on low-rank matrix completion problem (Case MC-S4: influence on larger memory size).

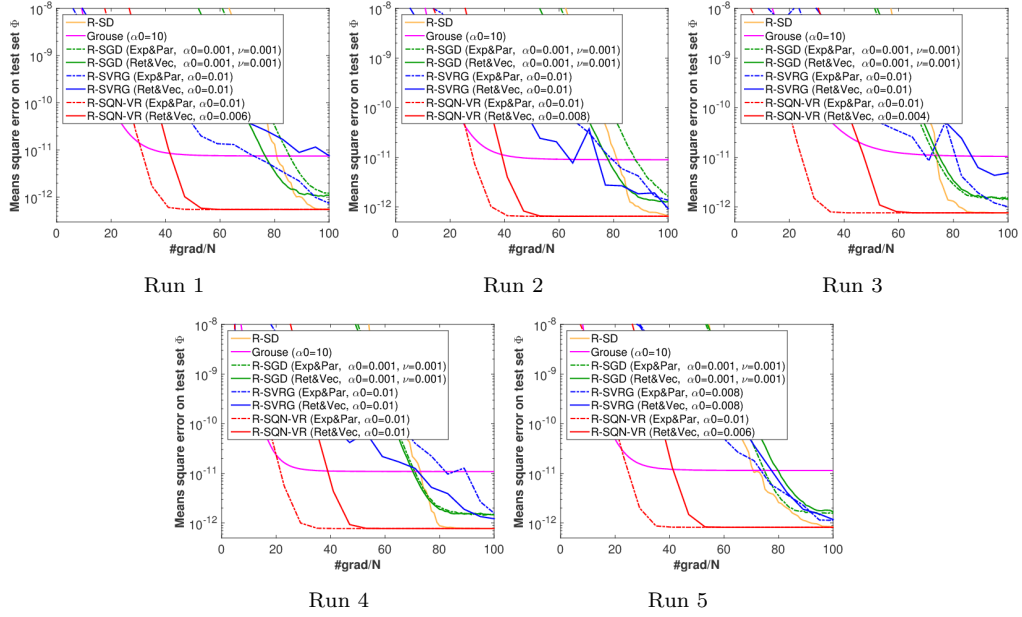


Figure A.7: Performance evaluations on low-rank matrix completion problem (Case MC-S5: influence on higher noise.).

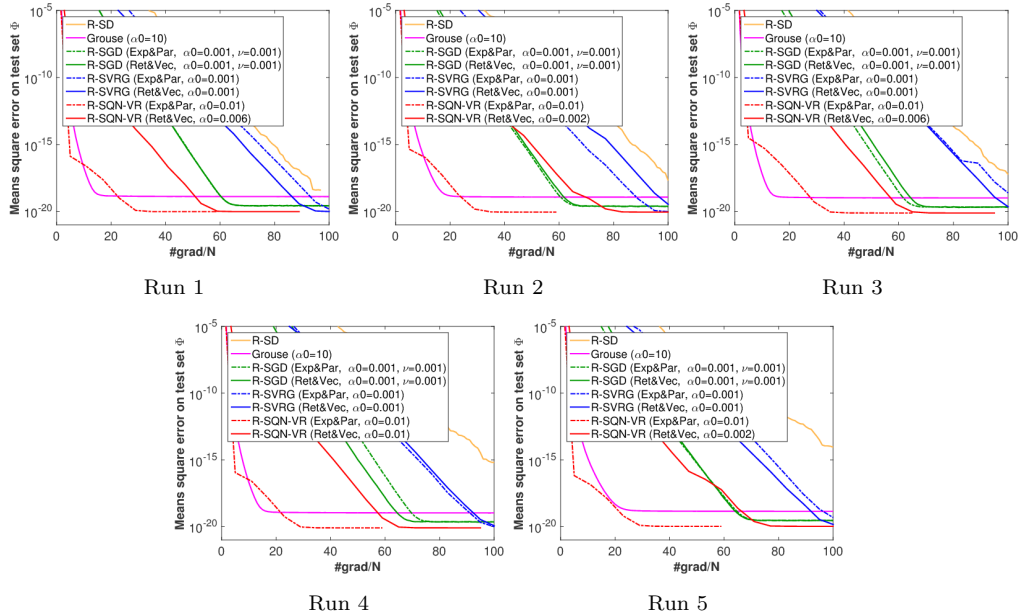


Figure A.8: Performance evaluations on low-rank matrix completion problem (Case MC-S6: influence on higher rank).

Case MC-S7: Comparison in terms of processing time. Finally, we attempt to confirm the performance of our proposed R-SQN-VR in terms of the processing time. Because all the algorithms are implemented by Matlab, it is difficult or might be meaningless to evaluate the actual processing speed. However, the comparison with R-SGD and R-SVRG, of which code structures are similar to that of R-SQN-VR, gives us useful insights to readers. Figure A.9 shows the results of the relationship between test MSE and the processing time [sec]. It should be noted that the result of “run 1” in each case is shown in this result.

First, it is noteworthy but not surprising that, compared with the results from the view-point of the number of gradients, where the cases of the retraction and the vector transport indicate slower convergences than the case of the exponential mapping and the parallel translation, these results show that the former cases indicate much faster than the latter cases. From these results, we confirm the advantage of the retraction and the vector transport. Next, the convergence speed of the SGD algorithm is much faster than others at the beginning because its process is much lighter than that of others. However, similarly to the results earlier, it converges to higher MSEs. Finally, the most remarkable thing in this result is that, in comparison of R-SQN-VR with R-SVRG, R-SQN-VR still gives superior performance than R-SVRG in terms of the processing time although R-SQN-VR requires one additional vector transport of a gradient in each inner iteration and L vector transports of the curvature pairs at every outer epoch. Consequently, we have confirmed the effectiveness of the proposed R-SQN-VR.

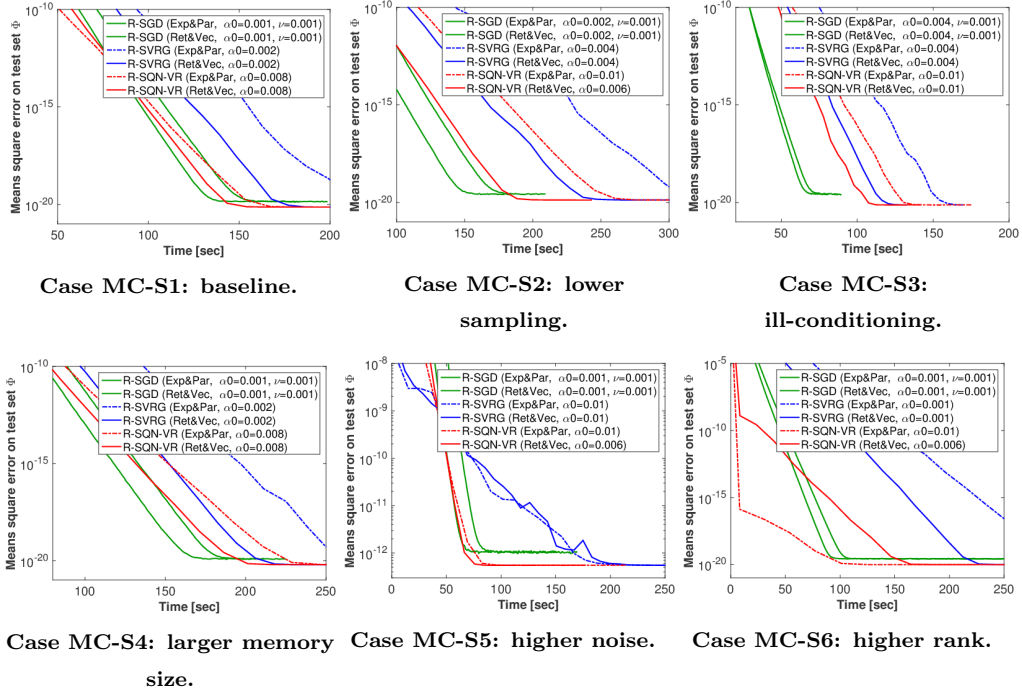


Figure A.9: Performance evaluations on low-rank matrix completion problem (Case MC-S7: comparison in terms of processing time).

F.3 Matrix completion problem on MovieLens 1M dataset

Figures A.10 and A.11 show the results of the cases of $r = 10$ (**MC-R1: lower rank**) and $r = 20$ (**MC-R2: higher rank**). They show the convergence plots of the training error on Ω , the test error on Φ and the norm of the gradient for all the five runs when rank $r = 10$ and $r = 20$, respectively. They show that the proposed R-SQN-VR gives a good performance on other algorithms, especially, when the rank is larger, i.e., $r = 20$.

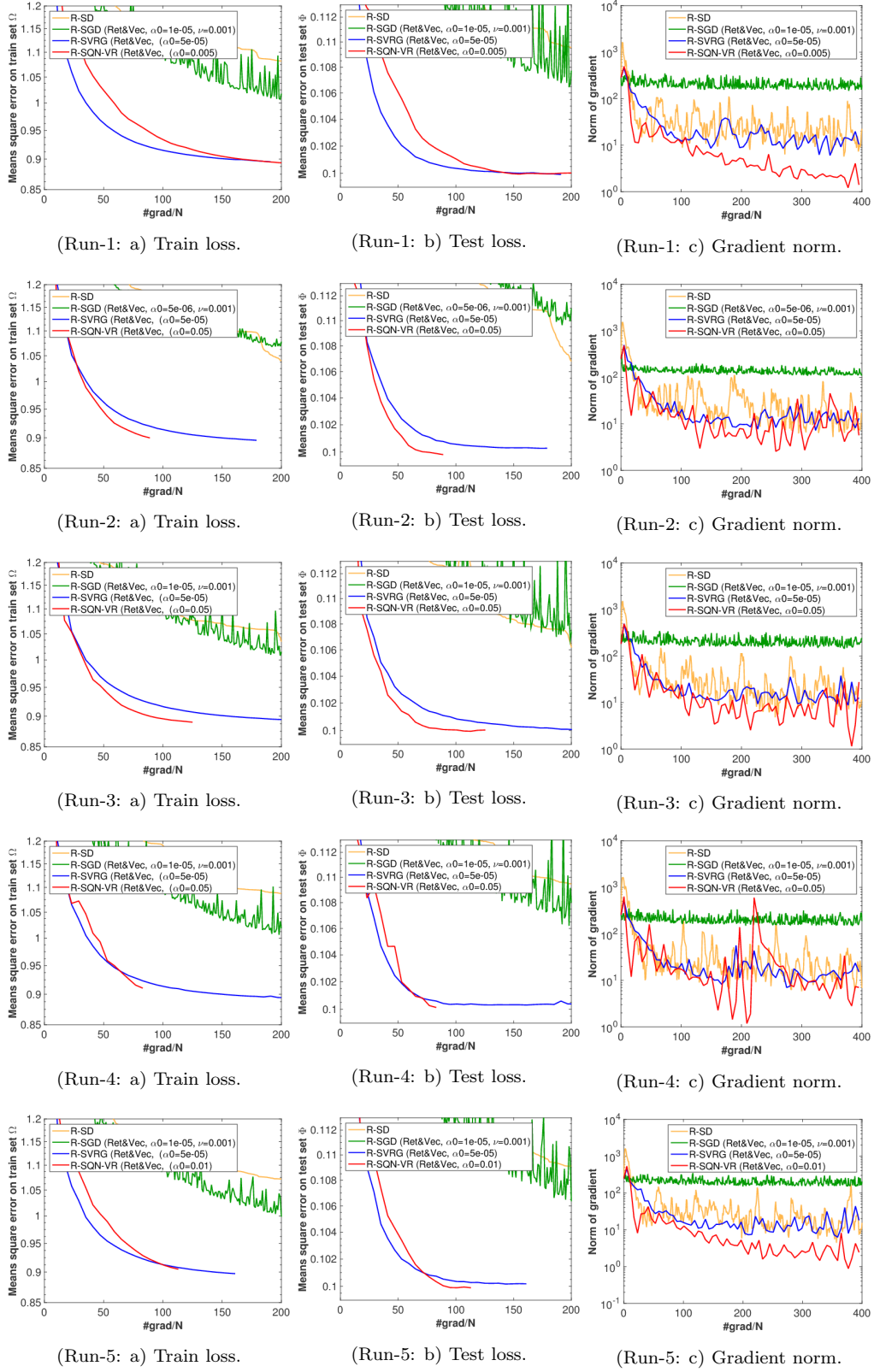


Figure A.10: Performance evaluations on low-rank matrix completion problem (MC-R1: lower rank).

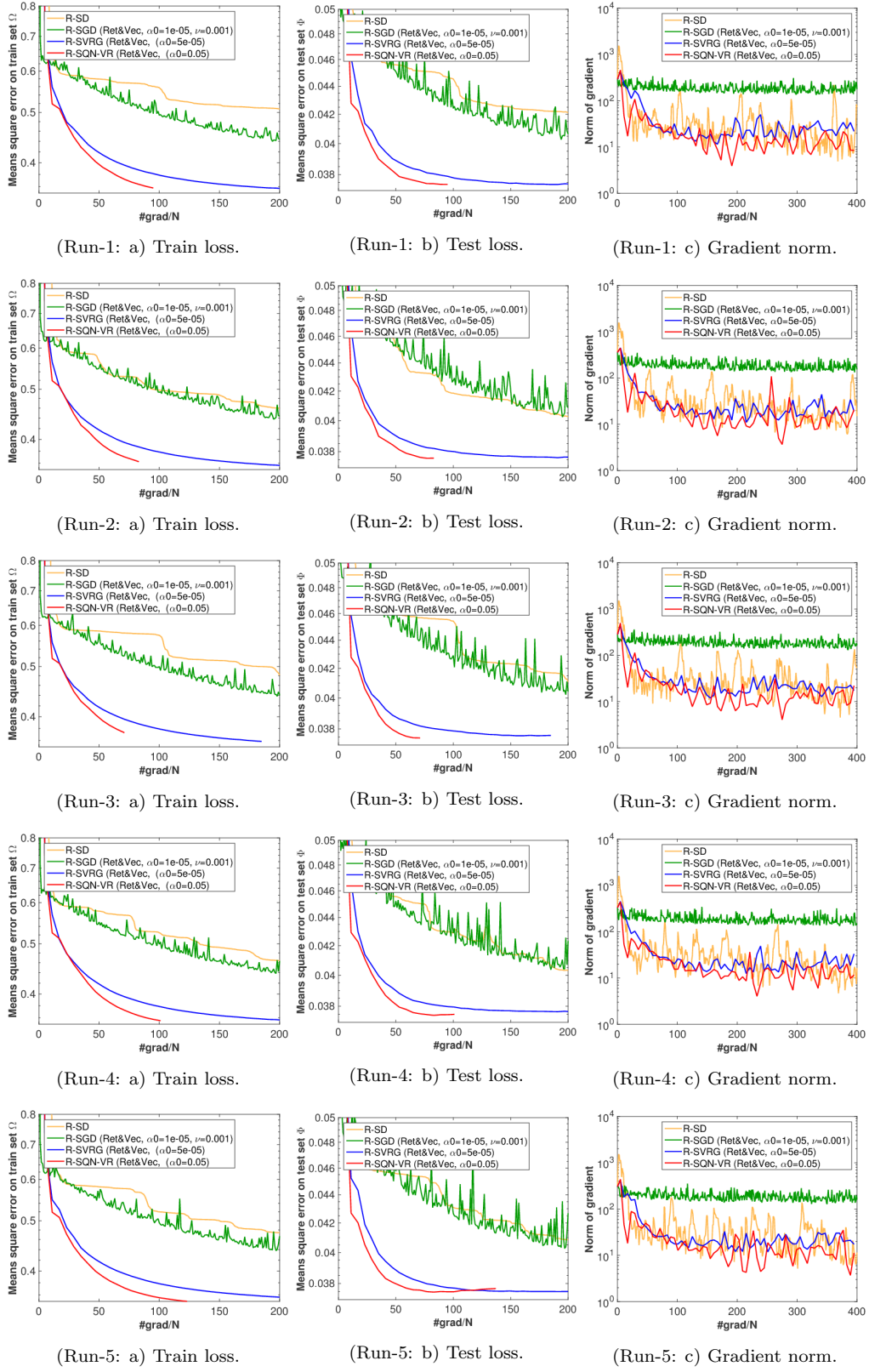


Figure A.11: Performance evaluations on low-rank matrix completion problem (**MC-R2: higher rank**).